2018

ANOMALY INFERENCE BASED ON HETEROGENEOUS DATA SOURCES IN AN ELECTRICAL DISTRIBUTION SYSTEM

Yachen Tang
Michigan Technological University, yachent@mtu.edu

ANOMALY INFERENCE BASED ON HETEROGENEOUS DATA SOURCES IN AN ELECTRICAL DISTRIBUTION SYSTEM

By

Yachen Tang

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Computer Engineering

MICHIGAN TECHNOLOGICAL UNIVERSITY

2018

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Computer Engineering.

Department of Electrical and Computer Engineering

Dissertation Advisor:	*Dr. Chee-Wooi Ten*
Committee Member:	*Dr. Laura E. Brown*
Committee Member:	*Dr. Panayiotis Moutis*
Committee Member:	*Dr. Yu Cai*
Department Chair:	*Dr. Daniel R. Fuhrmann*

Dedication

To my dearest mother and father, Mrs. Nanping Shen and Mr. Shimin Tang.

Contents

List of Figures

List of Tables

Acknowledgments

When I was writing this dissertation, I agreed to do my best to organize it using the traditional guidelines and have a title, copyright, acknowledgments, and an abstract all before you ever get to the table of contents and the actual text. In terms of advocating for the readers, general engineers, this thesis requires special treatment for description language.

I began working on this project in the spring of 2015 and outlining this dissertation in the summer of 2018. A three-and-a-half-year journey I certainly could not have navigated without help. Lots of it. Family comes first. Without the love and support of my parents, I would not be the scholar I am today. Regarding the creation process, I would like to express my deepest appreciation to my advisor, Dr. Chee-Wooi Ten, for providing professional development opportunities, support, feedback, and advice instrumental to my development as a Ph.D. student. His commitment to teaching, guidance, and research is extraordinary, and I cannot overstate the value of his support. Without his guidance and support, this dissertation would not be what it is.

I also benefited from some great professors. Dr. Laura Brown provided criteria for improving the relevant algorithm and the accuracy of output results. Dr. Sumit

Paudyal helped me to perfect the descriptions of models and formulas also gave me some hints from the professional aspect of power system operation. Dr. Yu Cai provided lots of valuable suggestions about fundamentals, background, and the future development to convince the audience. I would be remiss if I did not thank Dr. Kevin Schneider for his patient help about providing related materials and guidance of simulation analysis. Their helpful advice improved the preciseness of this dissertation.

Thanks to several of my colleagues who answered my questions on a variety of topics, lent their ideas, and sent their feedback on my work throughout its development: Mr. Zilong Hu, Mr. Wensheng Sun, Mr. Zhiyuan Yang, and Mr. Shuaidong Zhao, the Ph.D. students at Electrical and Computer Engineering Department and Civil Engineering Department. The writing coaches from the Multiliteracies Center got the largely thankless job of reading all the rough drafts rife with flaws and riddled with errors that eventually become a finished dissertation.

I am also grateful to my department, the Electrical and Computer Engineering department of Michigan Technological University, for the support of initial-phase metering deployment project.

And last, I would like to thank Drs. Laura Brown, Panos Moutis, Yu Cai and Sumit Paudyal for becoming my defense committee members.

Abstract

Harnessing the heterogeneous datasets would improve system observability. While the current metering infrastructure in distribution network has been utilized for the operational purpose to tackle abnormal events, such as weather-related disturbance, the new normal we face today can be at a greater magnitude. Strengthening the interdependencies as well as incorporating new crowdsourced information can enhance operational aspects such as system reconfigurability under extreme conditions. Such resilience is crucial to the recovery of any catastrophic events. In this dissertation, it is focused on the anomaly of potential foul play within an electrical distribution system, both primary and secondary networks as well as its potential to relate to other feeders from other utilities. The distributed generation has been part of the smart grid mission, the addition can be prone to electronic manipulation.

This dissertation provides a comprehensive establishment in the emerging platform where the computing resources have been ubiquitous in the electrical distribution network. The topics covered in this thesis is wide-ranging where the anomaly inference includes load modeling and profile enhancement from other sources to infer of topological changes in the primary distribution network. While metering infrastructure has been the technological deployment to enable remote-controlled capability on the disconnectors, this scholarly contribution represents the critical knowledge

of new paradigm to address security-related issues, such as, irregularity (tampering by individuals) as well as potential malware (a large-scale form) that can massively manipulate the existing network control variables, resulting into large impact to the power grid.

Chapter 1

Introduction

The existing distribution power system has been restricted to meet all of the development requirements because of the commodification of electric power, the corresponding adjustment of energy policy, the improvement of electricity utilization requirement, and the increase of renewable energy sources. Therefore, the hypothesis of the smart grid as the future distribution power system was put forward. Through the rapid expansion of smart grid in various countries, people received higher service quality for many aspects which include electricity generation, power transmission, distribution, and utilization.

Fig. 1.1 demonstrates the contrast of the current structure and the future structure of the electrical distribution grid in the graph representation. The nodes connected

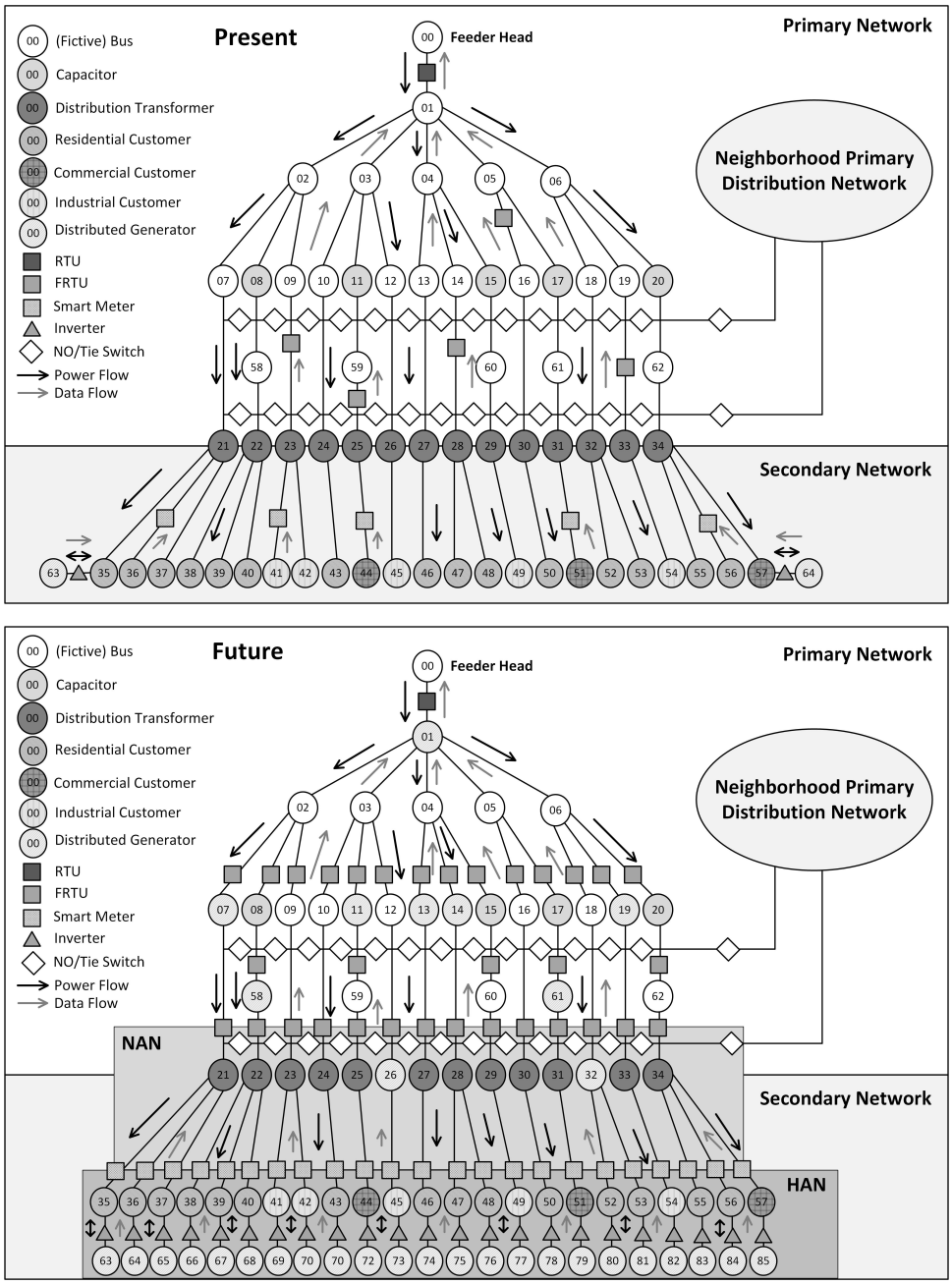

Figure 1.1: Contrast of the current structure and the future structure of electrical distribution grid in graph representation.

to distribution transformers in the secondary network are denoted as consumers (residential, commercial, and industrial). Only a few consumers are equipped with the

distributed generators, which include massive centralized plants, wind/water generators, small solar panels, and the growing roster of distributed energy resources as well, as the renewable energy sources and/or backup generators. The distributed generator from a renewable energy source should be operated connected through a power electronic based voltage converter, commonly in one stage that converts DC to AC. The DC to AC converter is called inverter and these distributed generators are as "inverter-interfaced" [1, 2]. The distributed generator can result in electricity flows in both directions through the inverter: from the distribution network to customers, and from customers with distributed generator back into the distribution network.

For power quality enhancement service, the grid-connected inverter of an inverter-interfaced unit has the similar structure as that of an active power filter. In general, due to the intermittency of solar irradiation or wind speed, the inverter cannot fully utilize its capacity and the idle capacity can be used to provide ancillary services [3, 4, 5, 6]. For inertia emulation service, the inverter-oriented distributed generators can mimic the performances of a synchronous generator for inertia emulation and power oscillation damping [5, 7, 8]. With more power from the inverter-interfaced units, the absolute reserve power from synchronous generators and the grid inertia constant are smaller due to the lower rated power of the rotating synchronous generators. As a result, the frequency response shows faster behavior with higher maximum frequency deviations when more distributed generation units are employed. The controllers designed to regulate the performance of the distributed generation units participate

also in improving the voltage stability in the system [1, 2].

The power supply in the distribution system is no longer a single "waterfall" mode, with electricity flowing from the substation in one direction towards customers. Current power system aspires to enable renewables penetration at a higher level by developing transformational grid control methods that optimize the use of flexible load and distributed energy resources. The renewable energy generated from distributed generators may be used onsite, or some or all of it may be exported to the distribution network. In addition, the traditional power stations in a power system could be a mix of energy sources, including renewable and conventional sources in the future.

The data flow direction is always from the users' end side to the feeder head. In current stage, despite installations of smart meters and supervisory control and data acquisition (SCADA) related recording devices, such as the remote terminal unit (RTU), and feeder remote terminal unit (FRTU), are increasing rapidly, the measurable area does not reach the ideal status that each load and bus should be installed with a real-time metering device to constantly observe the electric information. In the future stage, more sensors and more controllable devices should be deployed in the distribution system.

The diagram also shows the possibilities connect this network to the neighborhood primary distribution network in case of any fault occurrence around the system. The possibility of reconfiguring the topology should follow the rules that minimize the

power losses and customer interruptions. To estimate the power losses in operations, the power flow analysis is adopted. Since the distribution topology is radial, iterative techniques commonly utilized in the transmission network power-flow analysis, such as the conventional Gauss-Seidel (GS) and Newton Raphson (NR) method, are not converge for the distribution network. A ladder technique which consists of two parts: forward sweep and backward sweep, was the first solver examined to model three-phase unbalanced distribution systems. As shown in Fig. 1.2, all series elements in a distribution line segment, which include the overhead/underground lines, transformers, and regulators, can be represented as the aggregated component. With reference to Fig. 1.2, the fundamental equations for a ladder solver, i.e., Forward-Backwards sweeping (FBS) methods, are given by Eqs. 1.1 and 1.2, where $[A]$, $[B]$,$[c]$, and $[d]$ are generalized line matrices. The FBS methods require multiple iterations to converge even with linear or nonlinear load models. Under similar complexities, the number of iterations in a nonlinear load model will be clearly increased than in a linear model [9, 10, 11, 12, 13].

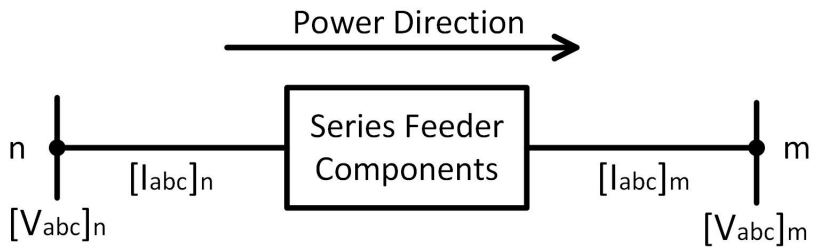

Figure 1.2: Standard feeder series component model.

$$[V_{abc}]_m = [A] \cdot [V_{abc}]_n - [B] \cdot [I_{abc}]_m \tag{1.1}$$

$$[I_{abc}]_n = [c] \cdot [V_{abc}]_m - [d] \cdot [I_{abc}]_m \tag{1.2}$$

Even though there are insufficient smart recording devices to record each load's consumption, the load modeling of each load can be estimated according to the existing SCADA system and historical consumption weighting factors. For example, Fig. 1.3 demonstrates a sample distribution feeder with the SCADA system. The RTU connected with the feeder head is monitoring the whole feeder includes FRTU1 and FRTU2. Since the tie switches connected to the microgrid are open, this RTU cannot monitor the operation status in the microgrid, but another RTU connected with the neighborhood primary network substation is capable of recording the usage information in this area. The metering summation of FRTU1 and FRTU2 should be lower than the recording in RTU because of consumptions of other nodes following the feeder and the power losses. Since the recording frequency of RTU/FRTU is every 3 to 5 seconds while the smart meter is every 15 minutes, the consumption curves from the FRTU and the summation of smart meters are with different time

representations. The recording summation of smart meters in this branch is lower than FRTU1 is also due to the power losses and metering errors. For the load area connected with FRTU2, only one load equipped with a smart meter, approximate methods have been used to gauge the weighting factors of individual loads which can be studied from a survey or to estimate using allocation factors (AF) based on transformer ratings associated with a root node of the metering point. As shown in Eq.(1.3),

$$AF = \frac{kW/kVADemand}{kVA_{Total}} \tag{1.3}$$

the AF can be determined based on the three-phase kW or kVA demand and the total connected distribution transformer kVA. The kVA_{Total} represents the sum of the kVA rating of the distribution transformers [14].

The home area network (HAN) is the basic communication infrastructure in the consumer side [15, 16]. It refers to the network that measure, collect and analyze consumption information from advanced smart meters through various communication media, for the purpose of forwarding the data to the grid data center and achieving home monitoring. In each household, the home monitoring system, smart data collector, and smart appliances can be interconnected through HAN. The neighborhood

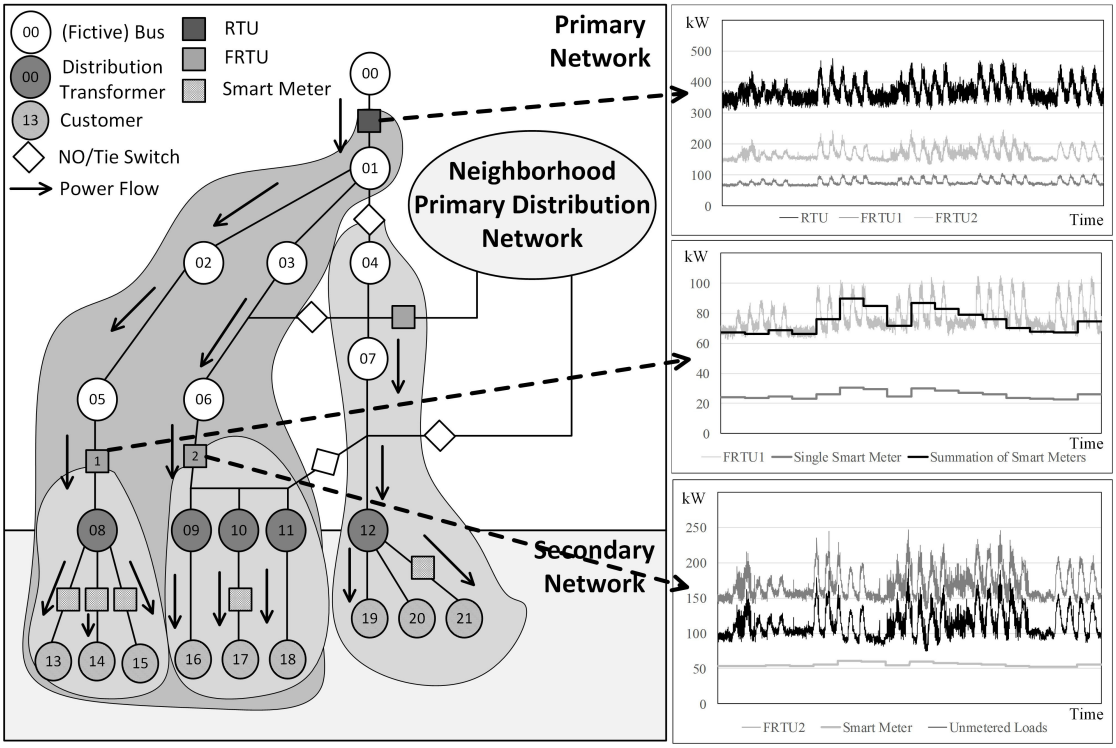

Figure 1.3: Distribution Feeder with SCADA system and corresponding consumption results.

area network (NAN) is the HAN complementary network that completes the distribution subpart of the smart grid. A NAN connects multiple HANs collectively for the purpose of accumulating consumption information from households (the HANs) in the neighborhood and delivering the data to the utility company eventually. Therefore, a precise and secure communication architecture and technology of the distribution grid are critical for the future system.

1.1 Grid Implementation Challenges

Once the future distribution grid has been deployed, it will incorporate numerous advanced technologies, such as real-time data exchange, data mining, and modeling, stable long-distance communication, cybersecurity, etc., to deal issues prevailing with conventional electric systems. The direct benefits of the future distribution system can be summarized as [17, 18, 19, 20]:

† The grid will be capable of accommodating wide various kinds of distributed generators safely and seamlessly as mentioned before.

† Consumers can monitor and manage the "smart" devices, choose the working and service time, avoid the power usage peak time, and pay bills according to the dynamic price.

† The grid is capable of providing different grades of service with varying power quality according to various pricing options and requirements.

† The disturbances and "faults" in the future distribution system can be detected, localized and prevented rather than simply react to them, and may act faster than operators ever could in resolving some emergency problems via the sufficient metering points.

† The protection devices and emergency measures in the future grid will address critical security issues from the natural hazards and also can provide a more safe and stable cyber environment.

However, the future distribution grid is still facing challenges in deployment [17, 18, 19, 20].

1.1.1 Inadequacies in Grid Infrastructure

Although the average advanced metering infrastructure (AMI) coverage in the US power utilities has risen, the rate of deploying AMI devices remains at a slow pace, modeling individual load consumption accurately within a distribution feeder has been challenging due to the limited metering points in the networks. The existing grid network is inadequate to accommodate the upcoming needs and requirements of real-time monitoring for all loads, transmission and detection of clean energy, and utilization of distributed generation, which may throw several challenges in design, erection, operation, and maintenance. Therefore, an alternative method utilizes a statistical approach to correlate estimated occupancy datasets associated with the power consumption can be developed to enhance the observability of a partially observable distribution feeder. This method is described in Chapters 3, 4, and 5.

1.1.2 Data Management

In the future, a complete power grid consists of an enormous quantum "smart" devices, such as smart meters, sensors, smart controllers, etc. The data from these devices and from other external sources like occupancy information, weather conditions, security alarms, etc. can help operators to enhance the safety and security of the system. An accurate analysis of archived data could help operators to avoid a breakdown or damage before the occurrence, improve the system operation, manage alarms, forecast demand, and price, etc. The data so collected is really big in volume. Voluminous data from these devices is not only difficult for collection and storage but also poses critical challenges in retrieval and handling. Database management is a vital issue in the distribution management system. The high volume of data may slow down the process of data collection, analysis and report generation.

1.1.3 Communication Issues

Even though there are numerous communication technologies have been deployed in the current power grid, the limited bandwidth, the limited exchange distance, the higher data loss, cyber attacks, etc. are still main issues in distribution system communication. There lacks a foolproof solution. Communication protocols do not

have a unique standard and are not well defined in the distribution grid network [21, 22]. The global system for mobile communication (GSM) and the general packet radio service (GPRS) have a coverage range of up to 10 km but they lack in data rates [21, 22, 23]. 3G requires costlier spectrum, whereas ZigBee is limited by coverage range of 30 to 50 meters only. Wired communication like power line communication (PLC) overcomes the issues of wireless communication but face the problem of interferences [22, 23]. The optical fiber is fast and secure but is a high price. Communication issues will lead to cybersecurity issues.

1.1.4 Cyber Security

The intelligent grid has been envisioned as a next-generation framework to modernize power grid and improve its efficiency and sustainability. Real-time monitoring has become a critical part of distribution network operation that enhances the control and automation capabilities as metering technologies evolve. The metering infrastructure has further extended from feeder head of a substation throughout the entire feeder loads. The AMI has become indispensable in a smart grid to support the real-time and reliable information exchange. However, computerizing the metering system also introduces numerous new vectors for malware and cyber attacks. Massive tampering of electrical metering is a notorious problem in the electric power system, which causes great economic losses and threatens the reliability of the power system. In smart grids,

smart meters may potentially be attacked or tampered to cause certain non-technical losses (NTLs) [24, 25]. It is challenging to identify malicious meters when there are a large number of users. Finding efficient measurements for detecting falsified consumption datasets and locating the tampered load have been active researchers in recent years. A graph-theoretic strategy for massive tampering detection based on profile comparison using reconfiguration switching schemes will be presented in Chapters 6 and 7.

As mentioned, connecting a power grid to a cyber network may trigger numerous vulnerabilities and intrusions in communication and data exchange of the system. Detecting, recognizing, and eliminating those loopholes before any security even safety threatens happens is essential. Three basic objectives of cybersecurity in future distribution grid are availability [26, 27], which refers to reliable and timely access to the database and other relative information, integrity, which includes the protection from an improper format, modification, and destruction of archived data, and confidentiality that refers to the security of data from unauthorized access. Cybersecurity is one of the substantial issues for operation since any single loophole has a potential threat to turn into the disaster for utilities and individuals involved with the grid. Since the distribution grid has a multilayer structure, each layer demands specific security concerns, requirements, and corresponding protections.

1.1.5 Privacy

Inadequacy in the vigilance of huge data handling poses a risk of potential consumer privacy. Safety and security of consumers information are of utmost concern. Breach of privacy of consumers information may occur as a consequence of any cyber threat or lack of proper policy as well. To maintain the faith of consumers, their privacy must be kept intact through cybersecurity as well as tough regulations. Hence complete assurance to maintain consumer's trust is required for acceptance of the real-time monitoring system in the future [28, 29, 30, 31].

1.2 "Abnormality" and "Anomaly"

1.2.1 Abnormality

The abnormality in the fundamental analysis of power system always refers to the electrical fault, which is caused by equipment failures such as transformers and rotating machines, human errors and environmental conditions. These abnormalities cause interruption to electric flows, equipment damages and even cause death.

The causes of power system abnormalities include weather conditions, which contains

lighting strikes, heavy rains, heavy winds, salt deposition on overhead lines and conductors, snow and ice accumulation on transmission lines, etc. These environmental conditions interrupt the power supply and also damage electrical installations. Equipment failure is another reason to generate abnormality. Various types of electrical equipment like generators, motors, transformers, reactors, switching devices, etc. causes short circuit faults due to malfunctioning, aging, insulation failure of cables and winding. These failures result in a high current to flow through the devices or equipment which further damages it. Electrical abnormalities are also caused due to unconscious human errors, such as selecting the improper rating of equipment or devices, forgetting metallic or electrical conducting parts after servicing or maintenance, switching the circuit while it is under servicing, etc. The abnormality might cause the damage to the power system [32, 33, 34].

The abnormality (fault) analysis conducted by reliability engineers involves locating faults, determining the cause, determining if protective equipment operated properly [35], and identifying what can be done to prevent the fault from occurring in the future. Under normal circumstances, a feeder monitoring system is installed to record voltage and current on every substation and every distribution feeder. Whenever the current or voltage exceeds a specified threshold value, the event is recorded along with pre and post-fault data to assist in the analysis of the event [34, 36]. The monitoring system consists of a mainframe server to archive data, a remote terminal unit (RTU) located in each substation and a feeder remote terminal unit (FRTU) at each feeder

to record the data.

A software package that allows the user to view waveforms associated with fault events as well as graphs associated with periodic data also installed in the system. Several automated reports are available in the software to view event data in various formats as well as provide diagnostic data on the health of the RTU and FRTU. The software also provides information on the operation of the substation capacitor banks which is very helpful in analyzing customer power quality problems.

In the power quality, abnormalities can include low power factor, voltage variations, frequency variations, and surges. When the power delivered to a system does not match what is expected, equipment can malfunction, prematurely fail, or not work at all.

1.2.2 Anomaly

Different than abnormality, the concept of the anomaly was promoted following the development of the smart grid. The operational and control center is the critical infrastructure of the electric power system. The restructuring of the power industry has transformed its operation from a centralized model to a coordinated decentralized decision-making model. In order to achieve the new model, the electricity management network, the corporate network, and the use of information technology

produce vulnerabilities and expose the electricity cyberinfrastructure to securities threats [37, 38].

Through the development of AMI, the supervisory control and data acquisition (SCADA) and energy management systems (EMS) as a part of modern power control system, play a vital role to monitor and detect the safety, reliability, and protective functions of the power grid. However, these systems, that were designed to maximize functionality with little attention paid to security, represent the potential vulnerability to service disruption or operational data manipulation that could result in public safety concerns. The anomaly in data and network layer of the modern power system refers to the unusual measurements with highly differing from old observed patterns that caused by network attacks, malicious tampering, viruses. It might not damage the power system. The anomaly frequently occurred through conscious sabotage from attackers and the main motivation is related to economic expenses.

1.3 Issues Associated with Anomalies

1.3.1 Malware

As the crucial part of the modern power grid, SCADA has a complex large-scale architecture, operate with real-time data exchange, and connect to the IP-based communication network. Therefore, keep secure and immune to malware attacks from SCADA during the normal operating is a massive challenge.

Malware attacks have evolved from the common internet worm and virus attacks to more precise attacks on target systems. While there have been significant damages by these internet worms and virus attacks, the present set of malware are designed to specifically steal information which is considered confidential, take control of systems for malicious purposes, create pathways (backdoors) through which other attacks can be launched or cause a complete breakdown of targeted infrastructures [39, 40, 41]. A typical example of such malware is Stuxnet [42]. Malware attacks on SCADA systems vary from mere invasive forms (e.g. to steal confidential information or to analyze the traffic of power supply by the system) to more invasive forms (e.g. to take control of the system or to cause a disruption in the normal functions of the systems) [43].

1.3.2 Fraudulent/Malicious Consumer

The fraudulent/malicious consumers is always a notorious problem in electric power systems over time and have serious implications for both utility companies and legitimate users. The fraudulent users tempering utility meters to disorganize or significantly reduce their billing information. In some situations, they have more energy used by rogue connections than the actual consumption, such as steal energy by connecting an extra cable to the power line without any energy meters. The fraudulent actions can be in the form of theft (meter tempering), stealing (illegal connections), billing irregularities and unpaid bills [44, 45].

The U.S., it is estimated that utility companies lose billions of dollars in revenue every year, while in developing countries, energy losses caused by malicious consumers can amount to 50% of the total energy delivered [46]. Energy fraud also leads to excessive energy consumption which may cause equipment malfunction or damage [47], and often enables criminal activities, such as the illegal production of substances.

1.4 Heterogeneous Data Sources in Distribution System

Average coverage of advanced metering infrastructure (AMI) in US power utilities has risen to 40.6 percent as of 2014 [48, 49]. Increasing the number of metering points improves load observability. A higher rate of AMI deployment occurred between 2010 and 2011 due to the Recovery Act Smart Grid Investment Grant (SGIG) program [50] in addition to increased utility investments. However, there is not sufficient data for the observability of the entire system, modeling each load consumption accurately within a distribution feeder has been challenging due to the limited metering points in the networks. For load model structure development, more load characteristics considered, a more flexible and sophisticated model can be constructed. Considering the possibilities of heterogeneous data sources that can be utilized and adopted in the distribution system will help operators and utilities to monitor and manage the grid well.

The application of heterogeneous data technology in the distribution network covers the entire procedure of data processing: from data acquisition and transmission to data identification, from data mining and processing to intelligent decision making, and finally to technology application and promotion.

1.4.1 Data Gathering Approaches

The current distribution system has sparse data collection points, narrow data transmission channels, and weak data processing capabilities. It is hard for managers to obtain valuable information and do efficient analysis through limited data sources. With the improvement of advanced power technologies, the distribution network data monitoring, acquisition, transmission, and management systems have been developed to provide a wealth of heterogeneous multi-source databases for distribution network operation and management.

Specifically, as shown in Fig. 1.4, the basic consumption information of the distribution grid as the fundamental information for the system operating and monitoring can be gathered from SCADA, the distribution management system (DMS), and the energy management system (EMS). The geographic information system (GIS) can provide the dynamic network topology and the geographic information while the operation status can be inferred and monitored according to the data generated in SCADA, DMS, and the distribution automation (DA) system.

Since the eventual task of the distribution system is to serve the social economy, economic development information of the society includes the investment in grid implementation is also one of the heterogeneous data sources that may directly affect

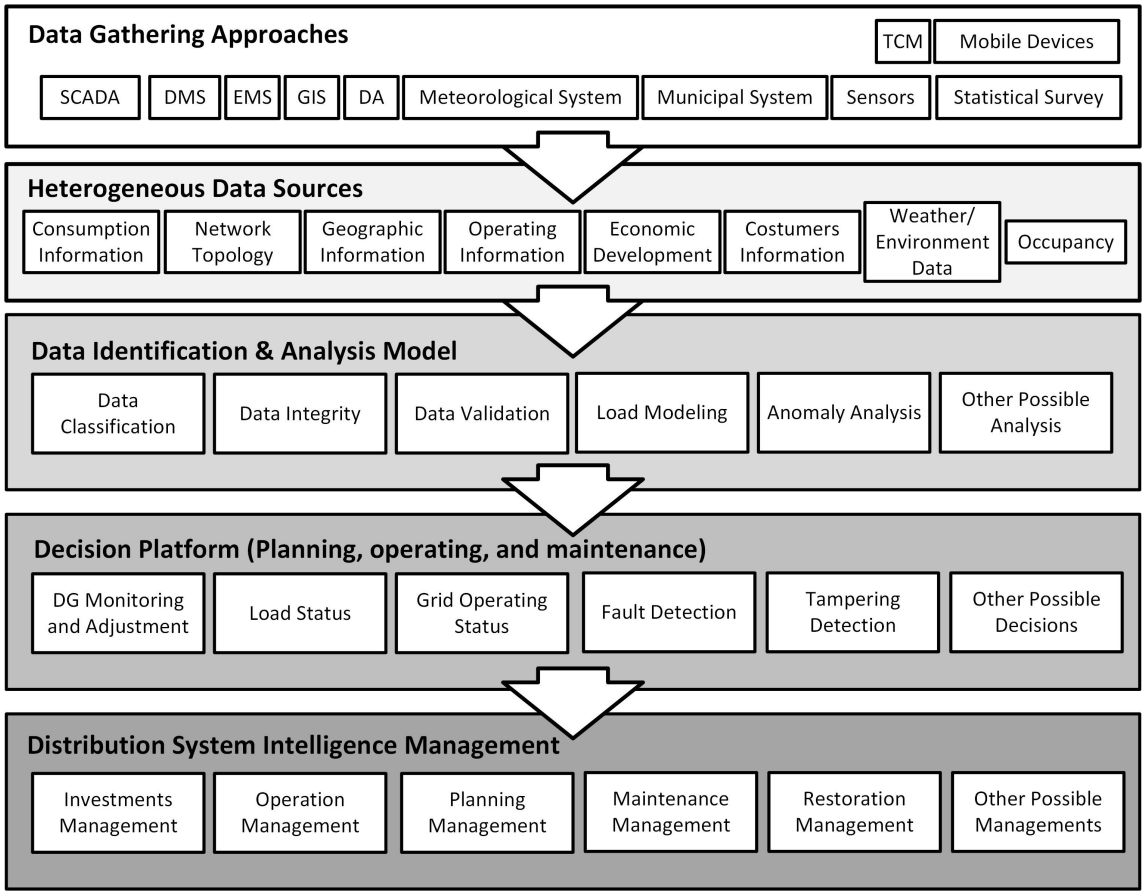

Figure 1.4: Flowchart of the application of heterogeneous data sources.

the development of the distribution network. This data source can be acquired from the municipal system.

Customers as the essential part of the power service, their load types, power quality satisfaction, power supply requirements, and complaints could also affect the operation and management of the whole system. Therefore, the customers' information can be treated with another heterogeneous data source gathered from the trouble call management (TCM) system [51, 52, 53].

The weather and environment data, including the temperature, humidity, carbon dioxide emissions, etc., could significantly impact power consumptions. Researchers have paid lots of attention on correlating the interrelationship between these two [54, 55, 56, 57, 58, 59, 60]. The extreme weather such as the hurricane, tsunami, tornado, and lightning could threaten the safety of the distribution system. Operators could adjust the system operation accordingly based on the weather and environment data collected from the GIS and meteorological system.

Because load consumption is closely related to human behaviors, different load models may be found in different time periods correlate with occupants movements and relative electricity activities. Even though the effects of occupancy are sometimes not noticeable. There are numerous examples such as large industrial facilities or agricultural facilities and automation that can be insensitive to the existence of occupants at the site, e.g., server farms, refineries and chemical plants, cold-storage facilities, irrigation pumps. In addition, in commercial and residential buildings, load error may often be closely related to weather. However, occupancy correlation is fundamentally related to power consumption. For example, customers adjust the settings of heating, ventilating and air conditioning (HVAC) system when they would be at home. As the most important section in heterogeneous data sources, occupancy data can be obtained from the human movement sensors, statistical survey methods, and the tracking of mobile devices. The detailed application and correlation between occupancy and power consumption will be illustrated in the following chapters.

1.4.2 Application of Heterogeneous Data

Fig. 1.4 demonstrates the flowchart of the application of heterogeneous data. Although heterogeneous data provides a variety of data sources to improve and enhance monitoring, operation, and management of distribution systems, not all of the input data can be directly applied for data analysis and modeling. Before the decision-making module, the data identification and analysis model is necessary. Different types of data must be systematically and effectively classified first. In the process of data collection, part of the data may come from manual reading and entry, such as statistical survey methods and TCM records. Manual input of large amounts of data is very likely to produce errors. Validation of data integrity, accuracy, and validity is the priority for data modeling and related analysis. The detected "bad" data should be eliminated. Commonly used identification and validation methods include extreme value theory [61, 62], residual correction method [63, 64], and pseudo-measurement detection method [65]. A detailed data validation process is presented in Chapters 4 and 5. Some new theories and techniques such as neural network methods, machine learning, and heuristic algorithms, which is applied in Chapters 4 and 5, are also utilized in data analysis.

Through the analysis of the characteristics of the customers' electricity usage behaviors, variations in the occupancy of a lumped load area, climate and environmental

impacts, dynamic changes in the power grid structures and other heterogeneous information, an adaptive power-consumption prediction/estimation model with higher accuracy could be established. This is also conducive to the realization of the power sources, including distributed generators (DG), and loads coordinated scheduling. The timely maintenance, power outage management, and restoration process according to the accurate detection of the potential fault in system operation could reduce power losses and improve users' service. Take remedial and penalties once a tampering behavior is detected. Eventually, the multi-sources heterogeneous data could help the distribution system to achieve the intelligence management in optimal investments planning, stable and safe operation, reasonable erection and implementation, timely and effective maintenance, etc.

1.5 Contributions

The main contribution of this thesis is to perform anomaly inference according to the existing database and heterogeneous data sources in an electrical distribution network. First, sensitivity analyses of occupancy and how it influences load consumptions under real conditions is performed. An agent-based model is then proposed to characterize occupant movement and electrical loads by generating datasets for this study. A statistical distribution with a heuristic regression model is then applied to correlate these two distinct properties to generate a time-dependent model. The

dynamic profiles of human movement and its load characteristics are established with parametric adjustments.

In addition, two real datasets, in terms of occupancy and power consumption (OP), are estimated according to a real case. Then, the sensitivity and correlation analysis are performed based on the regression method. Since model deviations from the a priori estimate of the regular pattern as a time-varying process with the curve fitting, the possible structural deviations in real-time demand are then considered to update the a priori regression model using new real-time estimates and observations obtained every demanded time duration.

The proposed correlation work offers an alternative to improve system observability of a distribution network with other interdependent networks. The applications of the proposed methods are to build a representative load profile that can be adapted to other unmetered load points, which could have tremendous value to having less frequent on-site data collections in a specific area of distribution network coverage as a result of labor savings.

As a distribution feeder has mixtures of load types that require specific treatment of load modeling. The extension of a load type to many, such as industrial, commercial, and residential, can be enhanced by relating to thousands of metered and unmetered loads within a feeder. Such a correlation with human movement would strengthen the load modeling. The crowdsourced information on an individual position from

the Internet about the real-time human movement in an area can be integrated for future work. Though, this enhancement might be limited to the availability of cellular devices an individual may have and carry with them.

With the trustable data source, the potential fraudulent load or clusters with malicious meters can be localized by combining the profile-based tampering detection method with switching operations of topology reconfiguration. A turnable threshold of anomaly will be applied to find the best balance between true and false alarm rates. The line section being detected does not need to be isolated. A graph-theoretic strategy for tampering detection based on profile comparison using reconfiguration switching schemes is presented. An intelligence searching algorithm is performed to reduce the searching time and iterations.

In the modern power system, the SCADA and the AMI network may potentially be attacked or tampered to cause the certain anomalous discrepancy. It is challenging to identify malicious devices or meters when there are a large number of users. The credibility of the measurements from the instruments will affect the safety and stability of the entire system. The alert system in meters will provide information support to operators or users according to different types of attacks or threats. However, some alarms will still affect the normal operation of the instruments, resulting in data losses, tampering, or trustworthiness. A massive tampering inference scheme

based on the variation of measurements from FRTUs and smart meters SMs is proposed. A probabilistic trust model based on alarm statistics to estimate possible or detectable data losses or tampering is also presented. Once the anomalous discrepancy exceeds the range of readings predicted by the trust model, an undetectable massive tampering could be inferred in the system.

1.6 Thesis Outline

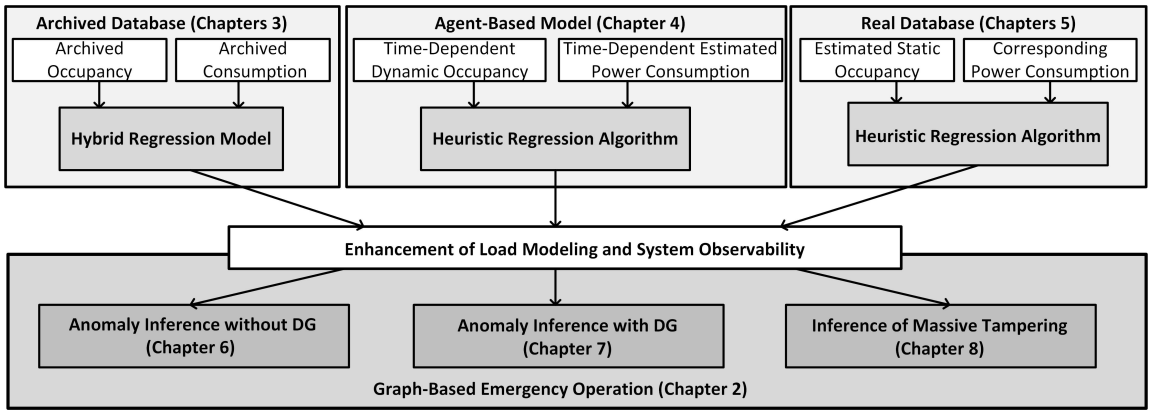

Figure 1.5: Overview of connection between Chapters in this dissertation.

Fig. 1.5 illustrates the overview of the connection between Chapters in this dissertation. Chapter 2 discusses the emergency operations in the distribution system according to graph modeling. The graph representation and model in this chapter is the essential part for anomaly detection and inference in Chapters 6, 7, and 8. In Chapter 3, a hybrid regression analysis is proposed to correlate the estimation of metered load profile with building occupancy in a distribution system that is conducted

based on actual datasets. Five possible regression candidates of load fitting functions for modeling the relationship between occupancy and energy consumption. The property regression models for load model in the distribution system are adjusted by correlating with data available from AMI, on-site reading, derivation of billing kWh information, or SCADA analog and digital measurements.

Since one distribution feeder has a variety of load types that require specific treatment of load modeling. By relating with thousands of metered and unmetered loads within a feeder, the extension of one load type to many, such as industrial, commercial, and residential, can be promoted. Human activities will also enhance the load modeling. The crowdsourced data about individual locations from the Internet that covers real-time human movement is integrated into Chapters 4 and 5.

In Chapter 4, a heuristic regression model is proposed to correlate the estimation of load profile with the number of occupants in a distribution system that is conducted based on an agent-based model. The adaptive parameterization of consumption aggregation from the number of customers as well as the determination of regression models from load profiles has demonstrated a promise in the proposed correlation framework. The proposed finite mixtures of regression models for load model in the distribution system are adjusted by correlating with data available from metering and/or survey recording results. The estimated occupancy can be gathered and synthesized from cellular devices around an area. Other technologies may be possible

but are not general enough to harness data in large scale.

In Chapter 5, a heterogeneous framework is utilized to enhance load models that incorporate the dynamics of occupants movements and how the existence patterns may influence the energy consumption. As each load may have a unique pattern that can be inferred from its activities, the proposed OP paired points can set as a statistical reference for inferring the consumption usages. The load profile enrichment can be established through the infusion of the historical energy consumption datasets. The proposed regression framework represents the observable dynamics of consumers energy consumption and movements which are used to correlate the patterns of two interdependencies. The adaptive parameterization of the occupancy-consumption aggregation has demonstrated the feasibility of the proposed correlation framework.

In Chapter 6, with improved system observability and sufficient data sources, an anomaly detection and localization technique using switching procedures based on the graph theory is proposed. The profile-based anomaly detection method is utilized to compare the consumption value displayed on the feeder head with the summation of all meters readings in its downstream. After localizing the anomaly load node, the comparison of consumption pattern for all meters connected with this node will be performed to find the tampered customer. In the process of localizing the anomaly load node, the distribution network is converted to a spanning tree to demonstrate the connection states in the topology and to avoid generating loops during switching

procedures, and then convert the spanning tree to incidence or adjacency matrix for future analysis. An anomaly localization algorithm is introduced to find the tampered load node. In addition, the cost consideration is mentioned to balance the benefit of localizing the anomaly point(s) within a large distribution topology.

In Chapter 7, a new approach to tampered node localization in full AMI using the concept of the graph theory is presented. The data-based tampering detection method is utilized to help indicate the existing of attacks or tampering activities in the system. An intelligence searching algorithm is performed to reduce the searching time and iterations. The consideration of subfeeders, clusters, or microgrids with distributed generators will reduce the number of combinations during the switching procedures that can accelerate the localization of the spot with the anomaly. In the process of localizing the tampered load node, the distribution network is converted to a graph only contains vertices and edges. The graph is demonstrated as incidence or adjacency matrix for the analysis. A tampered node localization algorithm within a distribution system is introduced.

Chapter 8 proposes a massive tampering inference scheme based on the variation of measurements from FRTUs and smart meters. A probabilistic trust model based on alarm statistics to estimate possible or detectable data losses or tampering is also presented. Once the anomalous discrepancy exceeds the range of readings predicted by the trust model, an undetectable massive tampering could be inferred in the system.

Finally, Chapter 9 summarizes the thesis work, discusses limitations, and outlines future research directions.

Chapter 2

Graph-Based Distribution Emergency Operation

2.1 Introduction

Emergency situations in the electrical distribution network are various and frequent, especially in system operation and management [66, 67]. The computerized management system for distribution operation plays a vital role in the emerging elastic framework in topological reconfiguration [68, 69, 70]. Electrical fault currents often cause protective relays to react to disturbances, resulting in large-scale power outages

that can affect thousands of customers. Locating fault segments or damaged components in a feeder between all border switches can be very time-consuming. The exchange of information between all entities, which include the existing indicators (such as fault indicators), trouble call tickets from customers, and field crew reports, can help to efficiently search for fault locations. By implementing a remote-controlled device on the distribution feeder, the search time can be reduced and achieve the temporary restore power to all customers in the "health" area by changing tie switches status. Eventually, the fault will be removed and restore power to all customers promptly. This can significantly improve overall system reliability in an emergency.

Due to the radial topology and unbalanced nature of the electrical distribution system, the topology can be represented by an incident and/or adjacency matrix (will be described in following chapters), where topological state and connectivity of interconnected feeders are critical to the application of the control center. The architecture of distributed SCADA/DMS is built on customer interaction management modeling, where the topology is significant for the structure used by the application. Topology is required in the management and operation of the power distribution system because dynamic geographic maps need to display [71] in the SCADA system and be utilized for operation analysis. These continuous updates are designed by distribution planning engineers to record the latest configuration of topology updates, including adding/removing/transmitting distribution transformers, new line segments or switches in the network. As the most common graphical representation platform,

GIS is often used to demonstrate incremental updates of topology information to ensure that the integrity and accuracy of these databases are up-to-date. These datasets describe all chart objects based on the data types, the measured value, and how it should be displayed on the map as well as symbols or drawing primitives. The static data and object data must be linked to the same SCADA device or metering element associated with the distribution node data, which is dynamically reflected by the system stimulus state.

Designing an unbalanced power flow module according to dynamic topology is critical in distribution engineering. Real-time evaluation using power flow modules will provide near real-world scenario simulations to predict the best results available for operational purposes. Computational or analytical methods should be developed to perform stringent operational type analysis, estimate system power loss, and meet the requirements of hypothetical analysis of large distribution systems connecting multiple substations. This is because the program will play an important role in determining unknown variables, such as the voltage and currents of all nodes, including all load nodes and distribution nodes (without loads), as well as the reactive power and angle of the generators (power injection nodes). This solution ensures a complete balance of the feeder when the power flow calculation converges. This is a very useful solution for determining the total active and reactive losses that can simulate a hypothetical scenario at a given time, and the solution with operational constraints [72, 73]. Different solutions provide multiple proposals to dispatchers in the system

control center to decide which tie switch they must turn on and off to temporarily restore the service to the "healthy" part of the area.

The main tasks of emergency operation in the distribution system is to assist the operator under faulted conditions and emergency operation to identify the fault type: permanent or temporary, find out the fault location based on fault indicators or trouble call tickets, perform fault isolation through changing the status of border tie switches, and achieving system or partial system recovery according to switching procedures (for RCSs) and/or field crew coordination (for non-RCSs). System operators and managers need to have a good understanding of the topology of the controlled network area in order to perform these tasks efficiently. However, topologies typically contain many components and various connection options. It is difficult to find the exact fault location accurately and quickly with limited information. Without the help of system automation, operators and managers must work with the training simulator to use these features proficiently and effectively.

2.2 Graph-Based Power Flow

The power distribution system is a low voltage system that maintains a radial topology or a temporary weak mesh network with a wide range of reactance and resistance values. The distribution system can include a large number of clusters or parts, such

as a microgrid. Due to the three-phase load imbalance and the presence of single-phase and/or two-phase loads in the lumped load region, there may be unbalanced loads in the distributed portion of the system. Phase frame approximation forward-backward sweep method has been proposed for power flow analysis in radial topologies and weak mesh networks. Fig. 2.1 demonstrate the process of the proposed graph-based power flow analysis. The up-to-date topology of a distribution system network can be constructed in GIS platform. A reduction model will be then applied to simplify the topology and delete the redundant nodes. The GIS data extraction process can obtain the geographical data from shapefile and generate the corresponding adjacency and/or incidence matrix to show the connection between nodes directly. Then, a data structure is constructed to illustrate corresponding data types, connections, quantifications, etc. for different components in the system. The detailed description of the data structure is described in the following section. Since the power flow approach is selected as the backward-forward sweep, the related "a", "b", "c", "d", "A", and "B" matrices can be generated from the data structure. Combining with the latest topology and the required data for each component, the backward-forward sweep can be performed to estimate the voltage and current for each node.

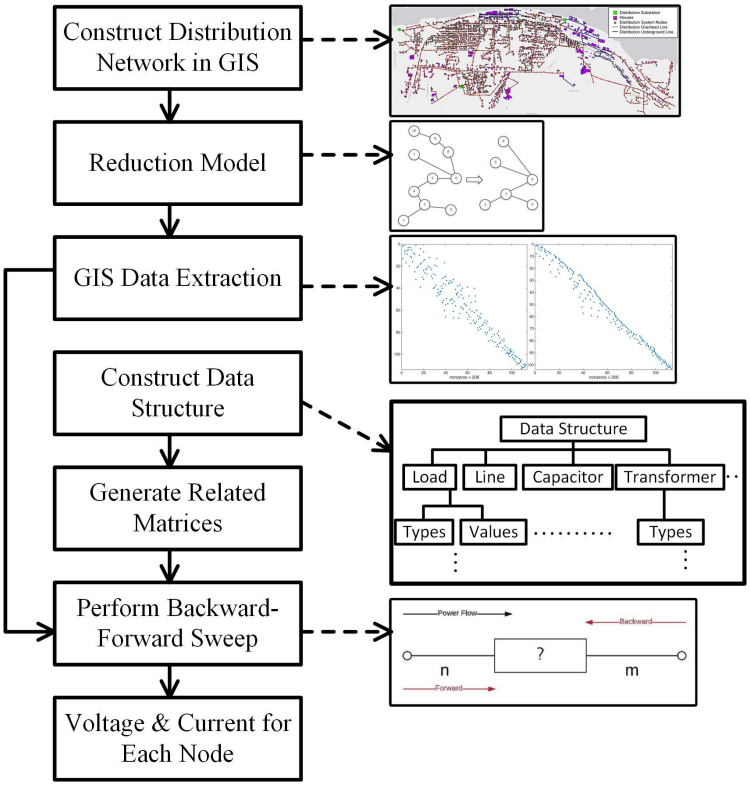

Figure 2.1: Flow chart of the graph based power flow analysis.

2.2.1 Geospatial and Topological Data Establishment

Extracting data from GIS is the process of retrieving demand data from the original shapefile and transforming the information into the required data form for further analysis and processing. Extracting geographic data from GIS is based on data attribute values, spatial extents, and geographic features. Using the extraction tool in GIS, the spatial reference can be used to extract the attribute information of the selected layer in the specified area to the desired format. New attributes or subsets of features can be generated for analysis.

Since the adjacency and incidence matrix can display the straightforward interrelationship between vertices and edges, the matrix is employed to store and demonstrate the current connection state of the network topology. A distribution feeder typically consists of hundreds of electrical elements in the network. The following enumerates the essentials of typical distribution components in a feeder [14]. Furthermore, this process converts the topology to the mathematic model which is machine recognizable. The further relevant algorithm and analysis can be performed based on the matrix[74].

1. Line segments (edges)

2. Distribution transformers (edges)

3. Switches (edges)

4. Voltage regulators (edges)

5. Loads (vertices)

6. Capacitors (vertices)

7. Distributed energy resources (vertices)

2.2.2 Reduction Model

In graph transformation and topological analysis, the reduction model in graph theory is used to simplify the work process and reduce repeated operations by eliminating redundant processes [75, 76, 77]. These processes can be blocks or clusters in a large-scale topology or can be vertices and edges in the graph. Similarly, power flow applications are performed on an equivalent graphical representation of a distributed network, which is typically very complex and involves a large number of edges and vertices. Depending on economic considerations and real geographic conditions, there may be multiple distribution nodes without any load in the deployment of actual feeders. These nodes are considered redundant nodes and may cause double counting during power flow analysis. The reduction function is used to eliminate redundant nodes and generate a relatively explicit topology.

2.2.3 Define Data Structure

A realistic distribution system is complex and contains multiple components. Generally speaking, the elements in a distribution system consisting of load, capacitor, regulator, overhead and/or underground lines, and transformers. The power injection sources or substations should be defined before the power flow analysis. The data

types and parameters for each component in a distribution system should be specified:

- † Load.
  - – Connection Types: delta or wye.
  - – Parameters: kVA, power factor, constant power, constant current, and constant impedance.
- † Capacitor.
  - – Connection Types: delta or wye.
  - – Parameters: kVAr and kV.
- † Regulator.
  - – Types: type A or type B.
  - – Connection Types: delta or wye.
  - – Parameter: tap.
- † Branch: Overhead or Underground.
  - – Overhead Parameters: phase conductor, neutral conductor, Cartesian coordinates of abcn.
  - – Underground: tape shield or concentric neutral.
  - – Tape Shield Parameters: cable data, thickness of tape shield, neutral data, and phase conductor diameter.

- Concentric Neutral Parameters: outside diameter, number of strands, phase conductor, neutral conductor, radius of the circle passing through the center of the strands, Cartesian coordinates of abcn.

† Transformers

- Types: delta grounded wye, ungrounded wye delta, open wye open delta, ground wye wye, delta delta, and open delta delta.

- Parameters for Each Type of Transformers: kVLL high, kVLN low, kVA, Z.

Since the forward-backward sweep approach is selected as the power flow method, some conversions need to extract and combine all relative information to generate the required matrix a, b, c, d, A, and B. The structure matrix is designed to display the connection states and all of its eigenvalues to be equal to one to ensure its invertibility. Then, the data structure with corresponding characteristics and parameters of each node and edge are incorporating.

2.2.4 Distribution Power Flow Analysis

In the operation and management of power distribution systems, power flow analysis plays a vital role in automation, including fault isolation proposals, switch combinations and service recovery in network reconfiguration. A large-scale distribution system has a complex topology and often update the connection status of tie switches, including load balancing and emergency operations. Therefore, the dynamic data structure allows dynamic topology reconfiguration, isolation, and restoration in the event of a failure, which is necessary for power flow analysis. Automation in modern power systems can handle these complex operations that require frequent topology changes (topology reconfiguration) in the power distribution system, which requires a dynamic topology logger based on a well-defined data structures[78, 79].

The forward-backward sweep approach is based on the ladder network theory. In the backward sweep, this approach calculates the distribution system bus voltages and adds the currents for each section. In the forward sweep, this method corrects the substation voltage on the feeder head with the initialized voltage value, the bus voltages in the system are calculated again in this iteration. No topological reconfiguration allowed during this backward and forward process and the approach meet the requirements of circuit laws. The convergence process will consist of multiple iterations and all bus voltages are computed twice in each iteration. After the first

forward and backward sweeps, the updated load voltages are calculated based on the most recent currents. The forward-backward sweep continues until the mismatch of the load voltages results between the new and the previous iteration is within the predefined tolerance.

2.3 Emergency Operation

The outage management system is one of the main components of the DMS, which is triggered by the occurrence of the fault(s). The existing fault localization technology is mainly realized by the fault indicator and the trouble call tickets form customers. A fault indicator is used to assist the management and control system to locate faulty components on the electrical distribution system. It usually equipped with a remote-controllable switch. In the fault localization analysis, the fault indicator only considered "Yes" or "No". Through the application of fault indicators, the search time required for locating faults can be significantly reduced. The goal of the fault segment location is to minimize the interruption of power delivery and promote power recovery in a timely manner, which will increase the reliability of the entire system. A fault indicator in the automation of the SCADA network provides clues that the fault location may fall within a certain area. This will help the field crew in the area to narrow down the faulted area by controlling the non-remote-controlled switches and determine the exact location eventually. The unaffected customers could recover

service by temporarily connecting the NO switch from another feeder.

The goal of fault isolation is to separate the faulted area, segment, or component of the network from the fault-free part. Isolation should ensure that the number of fault-free devices connected to the fault segment is minimized. In the graphical representation of the power distribution network, after the fault zone is located, the boundary remote-controlled switch connected to the zone is opened to isolate the faulty segment from the faultless zone.

The remote-controlled switches with fault indicators can localize and isolate an area first, the field crew change the states of non-remote-controlled switches and coordinate with the control center to narrow down the specific fault spot in the faulted area. The crew management system should be able to handle large amounts of data and complexities in multiple power outages. The system can receive and record outage events information from field personnel. In addition, the database of the system provides information about the spatial mapping of the power distribution system and the feeder facilities, including topology information for the feeder configuration, feeder facilities, and customer data. When the repair is complete, a historical event database is created for the failure event. Eventually, the crew management system will assist the dispatcher and operator by tracking the status of the incident, the crew's resources, and the effective assignment of field staff and tools to the incident.

The power supply for some customers will be temporarily restored in the faulted area

promptly and the number of affected users will be reduced after the crew coordination process. After removing the fault, the tie switches and configuration of the topology should be back to the normal condition. The flow chart of emergency operations in the electrical distribution system is demonstrated in Fig. 2.2.

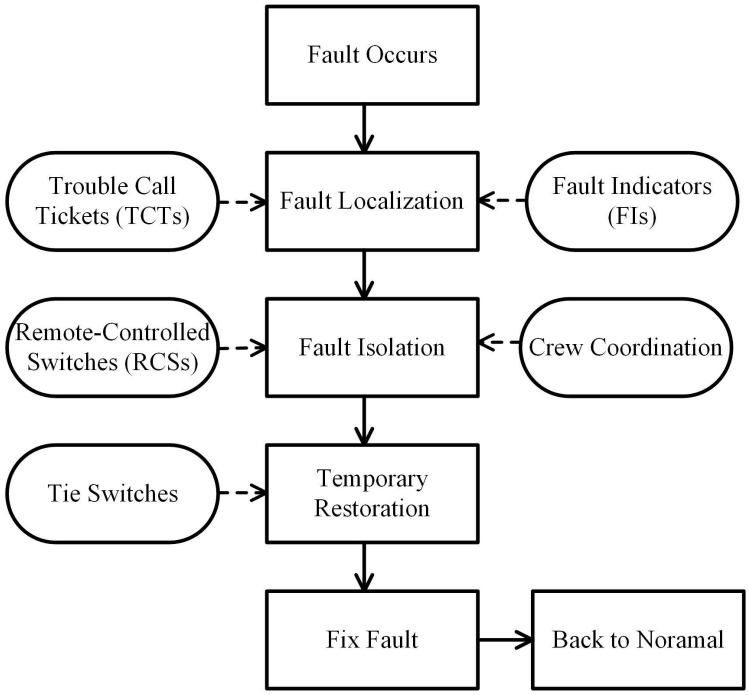

Figure 2.2: Flow chart of emergency operations in distribution system.

Fig. 2.3 illustrates an example scenario of emergency operations in a distribution system. A real power distribution system is much more complicated than this example. This simple scenario includes two feeders and includes multiple remote-controlled switches (equipped with fault indicators) and non-remote-controlled switches. Stage 0 demonstrates the normal condition of this system. The normally open (NO) switch break these two feeders.

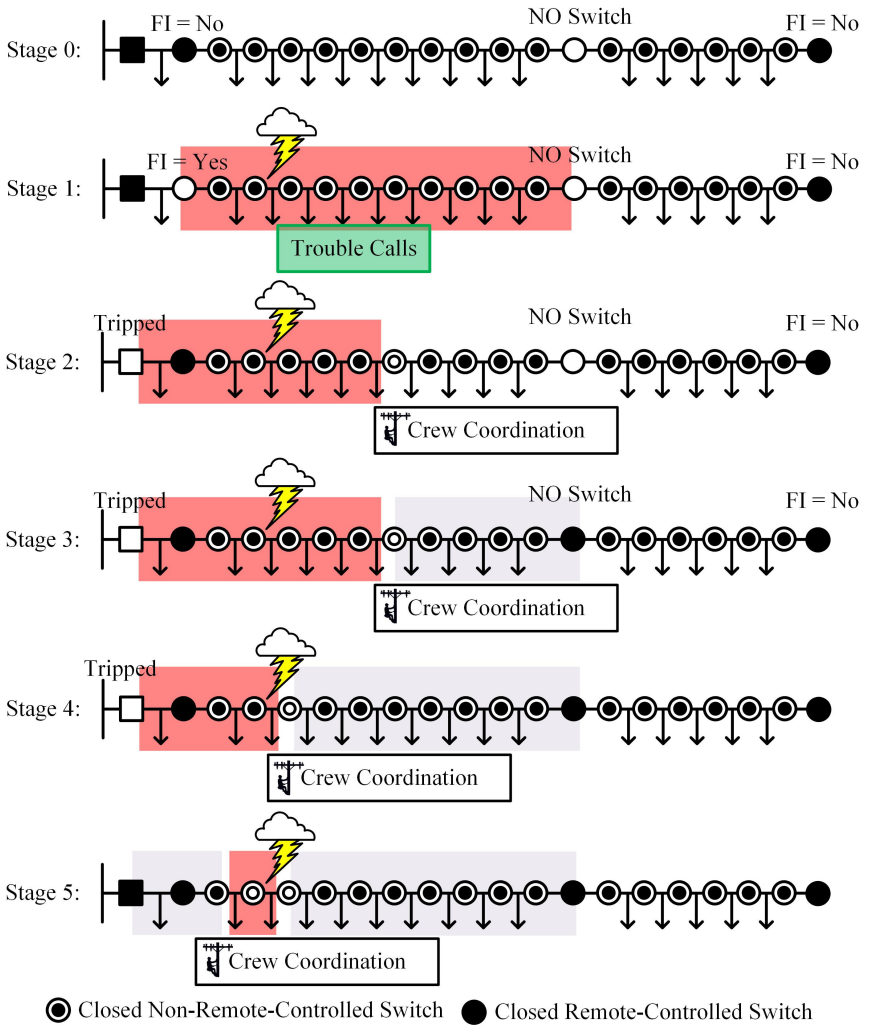

Figure 2.3: An example scenario of emergency operations in a distribution system.

Once a fault occurs in a feeder as shown in Stage 1, the fault indicator at feeder head shows there exists a fault and open the remote-controlled switch automatically to isolate the faulted segment. During this time, the operators in the control center could receive trouble calls from customers in the isolated area (shown with red color). In Stage 2, the breaker is tripped and the field crew will following the event logs, reports, or experience and coordinate with the control center to search the damaged

spot and narrow down the affected area. As shown in Stage 3, the NO switch is closed that the service of customers in the gray area is restored temporarily. Following the searching process, the affected area is reduced in Stage 4. Finally, the minimum faulted area is localized and isolated. After fixing the fault, the system can be back to the normal stage (from Stage 5 to Stage 0).

2.4 Emergency Operation Summary

With the development of SCADA system and the advanced metering infrastructure (AMI), distribution systems can be accurately analyzed and modeled. The SCADA metering device, FRTU and/or RTU is mainly the pole-mounted equipment in the primary distribution network, while the smart meters are equipped on the customer side in the secondary network. The customer billing center introduces a demand response based on the measurement feedback of the metering device to adjust peak and off-peak power usage and set dynamic pricing rules. Two-way communication between meters and suppliers in smart meters encourages customers to adjust their spending habits to better respond to dynamic electricity prices and manage their further assert. Although the AMI coverage of US electric utilities has risen, the deployment of real-time metering equipment is still slow, and because of the limited metering points in the network, accurately simulating individual load consumption within distribution feeders has been a challenge.

The distribution management system (DMS) plays a crucial role in the electrical distribution system. Once a fault or any other emergency situation occurs, the automation system in DMS will coordinate the information gathered from metering devices and/or indicators with the reports and experience of operators and field crew to deal with this problem. The emergency operations consist of fault localization, isolation, and system temporary restoration. After removing the fault, the system (including the operation states and topology) should be back to the original state. Real-time information exchange between the control center and the crew who are investigating field failures can speed up search time and greatly improve system reliability.

The possibility of topology reconfiguration can be within a feeder, between feeders associated with the same substation, or between feeders of adjacent substations. Each incremental update will be imported into SCADA to reflect the real scene. GIS is a versatile tool that helps planning engineers update outdated distributed network topologies. The GIS dataset can be extracted and converted to an incident or adjacency matrix. The incident or adjacency matrix can provide a visual representation to associate the connected elements in the distribution system. This will help us identify from the matrix where the possible adjustments and likelihood counts are expected to facilitate troubleshooting in the code. Although the actual situation is much more complicated than the virtual hypothesis, a similar assessment based on other combinations of distribution elements across the primary and secondary elements can be evaluated in the same way.

A resilient distribution system should not only be able to withstand failures but also be successful in more extreme situations, such as natural disasters or intentional violations by malicious consumers/organizations. Graphical modeling in the context of motivational state and reconfigurability will be the link for their future research and exploration.

Chapter 3

Enhancement of Distribution Load Modeling Using Statistical Hybrid Regression

3.1 Introduction

Enhancement of system observability with real-time monitoring in distribution networks has been promoted in recent years. Although efforts have been made to increase the number of new IP-based metering infrastructure with additional "smart" devices, the issues to accurately validate and establish the statistical distribution of unmetered

loads for a feeder has been at large dependent upon other already deployed IP-based devices in distribution networks. The IP-based metering infrastructure has been gradually improved over time by extending the single flow of information using automated metering reading (AMR) to bi-directional information flow using advanced metering infrastructure (AMI). Specifically, the applications based on the real-time metering information include state estimation, load allocation, probabilistic load flow, volt-VAr optimization (VVO), fault localization and system restoration for feeder reconfiguration and optimality. Modeling individual load profile can provide an enhancement of system observability to strengthen metering infrastructure and have a better understanding of the load behavior. The previous method requires metering individual loads in an entire feeder separately to determine the load models, which may not be practical in a short-term. Only the aggregated energy consumption information and data from metered loads are able to observe for the entire feeder or substation through smart meters deployed in the field.

For the energy consumption estimation of a building, the dependent variable is normally the electric utility usage, and the independent variables can include occupancy levels, weather statistics (e.g., average temperatures during test period or heating ventilation air conditioning (HVAC) condition), and utilization factors (e.g., area of structure, working days, power consumption rate of different utility unit). A comprehensive understanding of the load behaviors and characteristics can impact on the control and management to conserve energy. Predictably, the energy consumption of

a building during workdays is typically higher than that on the weekends and national or international events.

Most of the methods applied over the past years introduce statistical modeling based on the survey data for specific load types in the distribution system. A flexible regression analysis applied in this paper on different load shapes to present the relationship between occupancy and energy consumption. This analysis is a prioritization scheme that can provide an efficient method to enhance load modeling accuracy, system reliability, as well as distribution system state estimation (DSSE). Also, it can be applied to main building types after proving the feasibility of this study. One uncertain factor of occupancy was proposed to estimate the electricity consumption as well as to enhance load modeling.

3.2 Enhanced Load Modeling Framework

Regression analysis has been used for decades for dealing with the issues related to energy saving and optimization. A suitable model with the best fitting coefficient according to the trend and relativity of historical data can improve the accuracy of the prediction and estimation. The methodology of the analysis and modeling is presented in this section.

Fig. 3.1 shows the flowchart of the organization of this paper to present the framework for this work. Through the connection among all the blocks, the logic between each section is demonstrated. In this work, statistical energy consumption data and the number of customers in buildings are acquired as the response variables and donated as the statistical occupancy respectively. Dataset verification and preprocessing module justify the validity of input data. The statistical regression models are selected to implement the proposed hybrid regression analysis. Once regression models meet the requirements of validation, choose the best model to estimate the energy consumption within the corresponding time frame. Otherwise, the proposed adjustment of the dataset is applied to construct the combined regression model. The detailed descriptions for each module are discussed below.

3.2.1 Data Availability

The following types of data are commonly useful to a distribution system for load profiling purposes: (i) regular energy consumptions of individual customers, (ii) customer types, i.e., residential, commercial and industrial, and (iii) representative load profiles for each type of customers. For the collection of regular energy consumption data, despite the ongoing rollout of electricity "smart" meters, there remains a large portion of customers having no such capability deployed [50, 80, 81, 82]. One of the challenges of load modeling is a lack of the observation of measurement records for

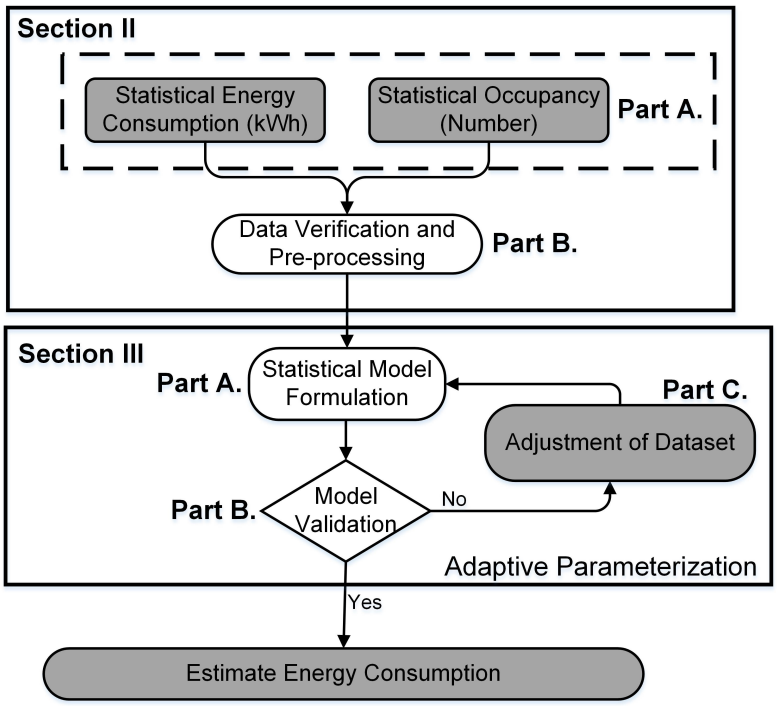

Figure 3.1: Organizational flowchart of the proposed correlation framework.

all customers.

For these customers, on-site readings of energy consumption are undertaken by crews on a regular basis (daily, weekly, or monthly). Furthermore, according to the bills and the price of electricity provided by the local power supplier or electricity utilities, the energy consumption during different time duration can be calculated. In addition, an image extraction method [83] can be utilized to obtain the real-time data from the eligible building as well. Also important is real-time information from SCADA. This feeder partition is useful in simplifying the regression analysis of system observability enhancement with a zonal approach.

The daily statistical survey result of occupancy for a specific load with a regular energy usage schedule can be obtained based on the records of the employee or the attendance of students. For the irregular loads, utilization of sensors and survey methods has been employed to collect statistical datasets. In transportation engineering, the study of human mobility can regularly infer traffic conditions from the cellular devices by tracking their positions [84, 85, 86].

In addition to the existing methods, a novel stochastic agent-based model of occupancy dynamics between loads with an arbitrary number of zones and occupants can be established for estimating occupants' movements. The maximum number of occupants in a building can be generated by the ratio of the available floor space to the average of an individual occupied area in this building. The dynamic occupancy of this building can be estimated according to the different attendance probabilities based on different time frames of each occupant who is called the agent in this model. Furthermore, the corresponding demand response (DR) of this building can be determined by GridLAB-D, which is a multi-agent simulation engine including advanced algorithms that can determine the simultaneous state of millions of independent devices, each of which can describe by differential and algebraic equations [87].

3.2.2 Data Verification and Preprocessing

The main purpose of the proposed model is to assure the accuracy and reliability of input data. This involves the process that eliminates the noises of input data caused by inconsistency data points of the empirical distribution. *accuracy*, *availability*, and *statistical model distribution* are considered as three data types of integrity constraints in the proposed model. Among all, the input data is assumed to have high accuracy and reliability in order to construct a distribution model. The records in input data should be continuous in a certain time period, otherwise, accurate short-term prediction cannot be guaranteed. In this integrity check module, some incomplete/missing and duplicated data points are detected. This operation guarantees, the input data from the metered loads corresponding to occupancy information before identifying appropriate models. Based on time-related factors such as the weather differs (temperature load) in different seasons, the mathematical relationship between energy consumption and occupancy can be varied and so as the consumption in weekdays, weekends, and special holidays. All input data are classified in this step by the time of day, the day of a week, the week of a month and the month of the year to reduce the error produced by time-related factors. The length of intervals selected for processing depends on prediction purpose (short-term or long-term prediction). For different classification levels, the parameters to determine the fittest model should change dynamically with different patterns.

3.3 Regression Models Incorporating Occupant and Metered Datasets

Human activities and movement patterns are inter-correlated with respect to their existence in a building as well as their contribution to energy consumption within the building. To answer how regression models can be built to identify the quantitative interdependency between human movement patterns and the number of occupants. The sensitivity of the patterns to the number of occupants is studied within a time-frame. As shown in Fig. 3.1, this section is a hybrid regression analysis to establish a load profile based on the metering and occupancy datasets.

3.3.1 Statistical Model Formulation

Even though the occupancy is dynamic, it is restricted by human activity patterns. The change follows the corresponding regularity of customer behaviors. Hence, regression analysis can be utilized to identify the quantitative interdependency between the independent energy consumption and dependent occupancy. Ideally, a simple linear model is probably the most direct approach to correlate the relationship between the variables. However, most actual data correlation may not exhibit a linear phenomenon, as a result, adaptive measures with hybrid functions are necessary to

capture the patterns precisely. Sometimes such cases vary that may accord with the trend of quadratic, cubic, or higher orders. The proposed method incorporates five most possible types of load fitting functions for modeling the relationship. The five regression models include linear, logarithmic, quadratic, cubic, and exponential models.

The regression coefficients can be calculated based on the least squares analysis and the unobserved random error of residual terms can be eliminated because calculations during the regression analysis follow a normal distribution of these error terms. The identification of the load modeling equations of different datasets can be based on the comparison of fitting coefficients from the five models mentioned before.

3.3.2 Model Validation

The module is a deterministic process to validate statistically from the preliminary formulation in the previous section. There are three tests to be evaluated which are based on (i) *R-squared test*, (ii) *F-test*, and (iii) *significance test*. These tests are to ensure consistencies of overall statistical distribution models that would best fit into the distribution model. As an important indicator in regression analysis, R-squared is the coefficient of determination, which not only indicates the goodness of fit but can also be interpreted as the amount of variation of the dependent variable explained by

the regression equation [88]. Generally, an R-squared of 0.75 indicates a reasonable correlation between energy consumption and the occupancy level [89]. F-values, which are created by the F-test [90], should be greater than interrelated critical values in the F-distribution table. The significance test, shown as "Sig.", shows the result of the statistical testing of significance. The value of this test result should be smaller than 0.05. In contrast, a meaning result of statistical testing should exhibit a significant difference between the F-test and significance test that would reflect on the observed data points and computed values.

The regression model of statistical curve fitting would require passing through the aforementioned three tests in order to form a reasonably justified hybrid model. The R-squared test must be executed after F-test and significant test. If any of the two fails, then R-squared will not be initiated. Different test model would have the minimum threshold value, e.g., the R-squared test with at least 0.75. However, we use at least 0.8 value in our study as the threshold of the R-squared test because it improves the fitness of the proposed model and avoids overfitting.

3.3.3 Adjustment of Dataset

This module will be applied only when the R-square test of the 10 given candidate models in model validation cannot meet the requirement. Withdraw the last 10

percent samples to format two new training sets and re-perform statistical model formulation. Repeat the iterative procedure until the model satisfies the criterion. Therefore, the proposed model may be a finite mixture of regression models.

3.4 Regression Analysis for a Case Study

This section is to validate the proposed regression analysis with the realistic load metering datasets of a campus building in Michigan Tech. The occupancy information is estimated based on the classroom schedule and students registration status using Fall 2015 semester datasets.

The test building was selected because all of the three circuits associated with that building have IP-based energy meters. This building is one of the high traffic building that has the most human movement that would be helpful for us to establish a statistical study using the proposed models. The building is one of the highest consumption loads across all campus buildings.

3.4.1 Time Windows of Metering and Occupancy Datasets

The time windows considered in this data gathering are based upon the consistent intervals that are captured for occupancy and metering information. As most buildings

follow the weekly routine and the weekly data pattern is typically chosen for regression analysis, the data is chosen based on the week interval during the test period. According to what we have proposed in this work, it is best to build a norm profile in the third week of the semester. The time windows of the Fall 2015 semester datasets are between September 28, 2015, and October 4, 2015. The number of enrollments reflected on these weeks has shown relatively stable since students tend to confirm the courses they wish are satisfied with the experience in the past two weeks. And it is unlikely to change due to the low refund rate if they do so later in the semester.

The data transferred from the devices to the building management system via the Internet link module (ILM) rather than wireless connection due to the higher data reliability. The occupancy information is mainly gathered based on the students registration records from the administration database. The detailed course schedule can be used to ensure the classroom location and its association with energy consumption with a particular building. Also, this estimation includes the number of faculty members, staffs, and graduate students who have their routine to the tested building. This number varies building to building that can be insensitive to its others. As weekdays and weekend may have statistical fluctuation, we only use five workdays data in this study. In addition, the work days should also be divided into work hours and non-work hours to improve the precision of the model. A unit time interval is selected for processing depends on the precision preference and the practical situation. The time duration is selected as 10 minutes in this case.

3.4.2 Establishing Statistical Models without Temperature Load Consideration

Figs. 3.2 and 3.3 illustrate the initial regression model of work hours and non-work hours sample data in the test building. The x-axis represents the number of occupants and the y-axis is the energy consumption (kWh). Five regression models are applied in this study based on Section III.A. The Department of Electrical and Computer Engineering is in this building and the total enrollments of students in fall 2015 were 511, which contains 319 undergraduate students and 192 graduate students. The normal work hours are from 9:00 am to 5:00 pm while the main courses hours concentrated on 10:00 am to 4:00 pm in that semester. The reason why the occupancy excesses the total number of enrollments is because there are 20 staffs and 52 faculties working here, the school of technology and IT department also share this building and some courses can be registered by other departments students.

Tables 3.1 and 3.2 are the summary of the initial validation. Each column shows the corresponding parameters to the five regression models. According to the validation criteria, R- squared values of all models in Fig. 3.3 is larger than or close to 0.75, the F-values are far greater than the critical values in the F-distribution table, and the Sig. values are all 0s which are smaller than 0.05. Each of the columns shows

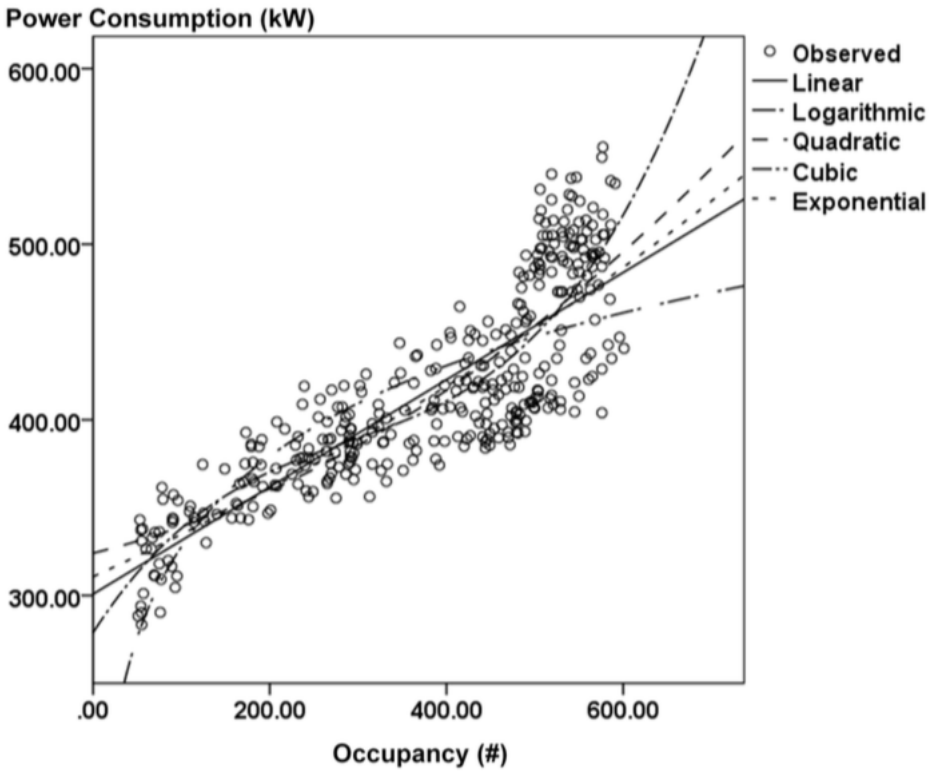

Figure 3.2: Model 1: Statistical regression models on test building.

consistently that they are more within their threshold values. However, all of the R-squared values in Fig. 3.2 are less than 0.75. The adjustment section will be applied in the work hours dataset to construct a hybrid model, which is shown in Fig. 3.4, to meet the requirements of the criteria. Table 3.3 is the validation summary of the hybrid model. The weight values of submodels in work hours model are approximate to 0.66 (208 points), 0.21 (65 points), and 0.13 (42 points) and the model types are quadratic, linear, and quadratic, respectively. The type of non-work hours model is cubic.

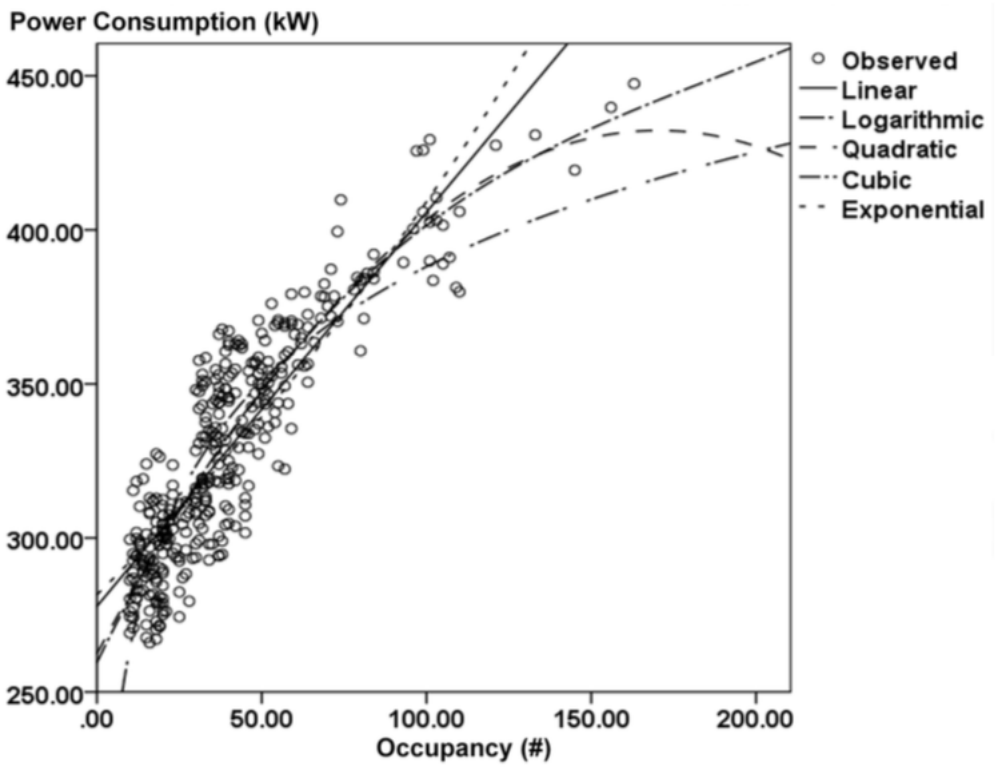

Figure 3.3: Initial regression models of non-work hours sample data in work days on the test load.

Table 3.1
Sample Data of Initial Validation Summary During the Work Hours on the Test Load.

Model	R-squared	F	Sig.
Linear	0.701	852.967	0
Logarithmic	0.618	586.650	0
Quadratic	0.712	448.168	0
Cubic	0.730	324.698	0
Exponential	0.733	994.813	0

Table 3.2
Sample Data of Initial Validation Summary During the Non-Work Hours on the Test Load.

Model	R-squared	F	Sig.
Linear	0.775	1216.247	0
Logarithmic	0.765	1148.546	0
Quadratic	0.805	727.834	0
Cubic	0.806	485.283	0
Exponential	0.744	1024.174	0

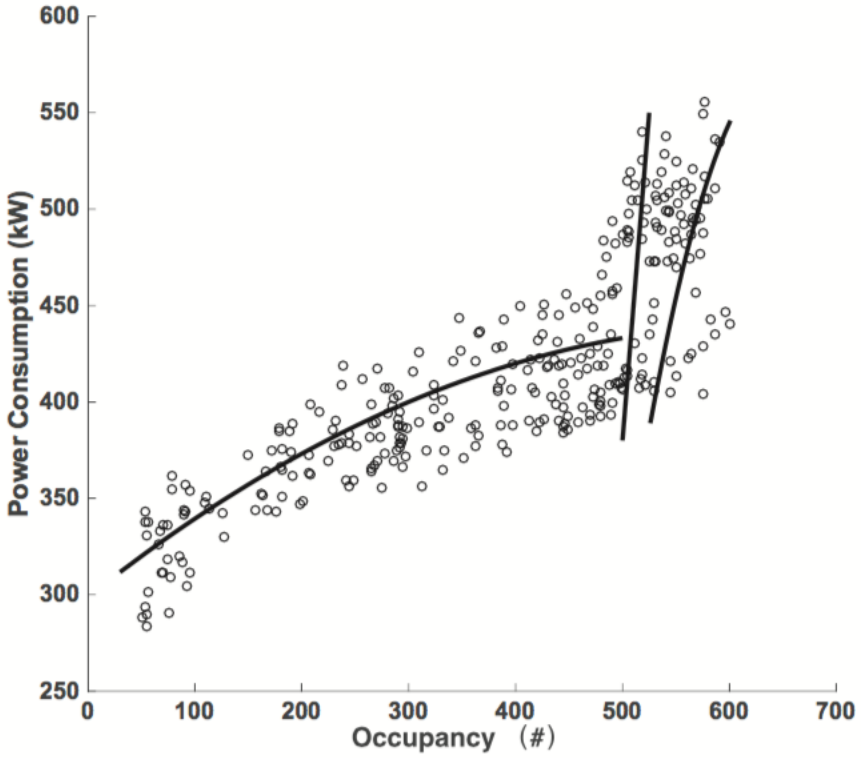

Figure 3.4: Hybrid regression model of work hours in work days on the test load.

3.4.3 Study Results

Tables 3.4 and 3.5 illustrate the error rate analysis between the overall regression model and the true value captured from a random week without consideration of

Table 3.3
Validation Summary of Hybrid Regression for the Test Load.

Model	R-squared	F	Sig.
Sub1	0.801	561.181	0
Sub2	0.833	833.998	0
Sub3	0.788	603.268	0

weather factor. Two pairs of the data which include the cumulative occupancy and the total energy consumption within different time durations were randomly selected. Among all, one was selected from the accumulation points of work hours while another one was in non-work hours.

Table 3.4
Error Rate Analysis of the Test Building in Work Hours.

Model	Occupancy (#)	Energy (kWh)	Error Rate (%)
Actual	263	361.63	-
Model: Cubic	263	372.54	3.016

Table 3.5
Error Rate Analysis of the Test Building in Non-Work Hours.

Model	Occupancy (#)	Energy (kWh)	Error Rate (%)
Actual	43	351.28	-
Model: Cubic	43	346.57	1.341

Tables 3.4 and 3.5 show the number of occupants and the observed value of energy consumption of a campus building, i.e., test building, are 263 and 361.63, 43 and 351.28 respectively. The error rates of regression models chosen from the proposed model are under 3%, which are in an acceptable range. The error includes the statistical error in data collection and the error might be caused by other factors such as the abnormal weather or special events. Predictably, more sample data (monthly or

quarterly, not yearly because the influence factors are distinct) and more uncertain factors considered can construct a more accurate regression model.

Chapter 4

Enhancement of Electrical Load Characterization

4.1 Introduction

Load modeling has been an important subject in understanding demand response and energy management [91]. In operational planning, load models play a crucial factor to estimate losses as well as to ensure voltage profiles of feeders fall within operating limits [92, 93]. All customers under one feeder represent thousands of electrical loads. Ideally, each load should be metered with an advanced metering infrastructure (AMI) device to report their energy consumption. Although the average AMI coverage in

the US power utilities has risen [48, 49, 81], modeling individual load consumption accurately within a distribution feeder has been challenging due to the limited metering points in the networks. Due to diversity and load variation, an ongoing survey on specific profiles of unmetered loads has been conducted constantly [94, 95]. Approximate methods have been used to gauge the weighting factors of individual loads which can be studied from a survey or to estimate them using allocation factors based on transformer ratings associated with a root node of the metering point [95].

Some utilities do not implement full smart meter deployment. Hence, their load models for each household are often updated regularly. Such labor-intensive statistics may not describe their consumption activities. A heterogeneous alternative to characterize their movements would establish a new profile of load models that can be incorporated. A load profile can be enriched with a balance of flexibility and a certain degree of complexity. An advanced algorithm can be established to discover the new load patterns based on consumers' dynamic behaviors. Consumption of electricity can be irregular since the occupants' activities may not always be uniforms, such as adjustment of HVAC, heat produced by customers, and their activities, e.g., lights and plugs; thus, an enhancement of electrical load modeling should be found to correlate with occupants movements and their physical existence and relevancy in their electricity consumptions. The common profile establishment is often a specific snapshot by a random survey to a household that cannot thoroughly represent other load

similarities as each load is unique [91]. However, characterization of consumers' physical existence to their electricity consumptions can be correlated and incorporated in the existing load model.

In transportation engineering, the study of human mobility can regularly infer traffic conditions from the cellular devices by tracking their positions [84, 85, 86, 96]. A data analytic correlates information based on a regression framework to model electrical usage [97, 98, 99]. From the recent studies [97, 100, 101, 102, 103, 104, 105], the occupants' behavioral data input can be synthesized to optimally curtail electrical loads. A framework can be established to enhance the dynamic behavior of occupants. Derived occupancy datasets and their energy usages are combined based on the data from the US Energy Information Agency (EIA) or the American Time Use Survey (ATUS). In addition, environmental parameters, such as temperature and humidity, are included in the studies to discover patterns of a given building [99, 104, 106, 107]. Sometimes, the effects of occupancy may not be noticeable. There are numerous examples such as large industrial/agricultural facilities and automation that can be insensitive to the existence of occupants at site, e.g., server farms, refineries and chemical plants, cold-storage facilities, irrigation pumps. In addition, in commercial and residential buildings load error may often be closely related to weather.

However, the occupancy-consumption correlation in a residential area can be obvious. For example, customers would adjust the settings of the heating, ventilating and

air conditioning (HVAC) system when they are at their residence. Prediction and utilization of energy usage at the appliance level can be estimated and improved [108, 109, 110, 111, 112]. Control management of local electricity usages is further investigated based on offline occupancy and demand response with improved methods [97, 101, 106, 113, 114, 115, 116].

Fundamentally, Fig. 4.1 shows the statistical correlation between occupants' movement and their energy consumption associated with metering devices. These loads are spatially correlated and associated to metered and unmetered zones. The four clusters in green represent the electrical loads with smart meters, which can be used to calibrate the proposed model directly. The three scenarios with different time indexes show diverse occupancy densities. As the ideal case of the statistical curves illustrates in Fig. 4.1, under normal circumstances, the power consumption in this area is expected to change at different time frames that are influenced by their existence closer to the metering points. Under the preliminary study, these two distinct dimensions of occupancy and electrical loads exhibit varying patterns that reveal their dynamics [117]. However, the load variation may not obviously be discerned.

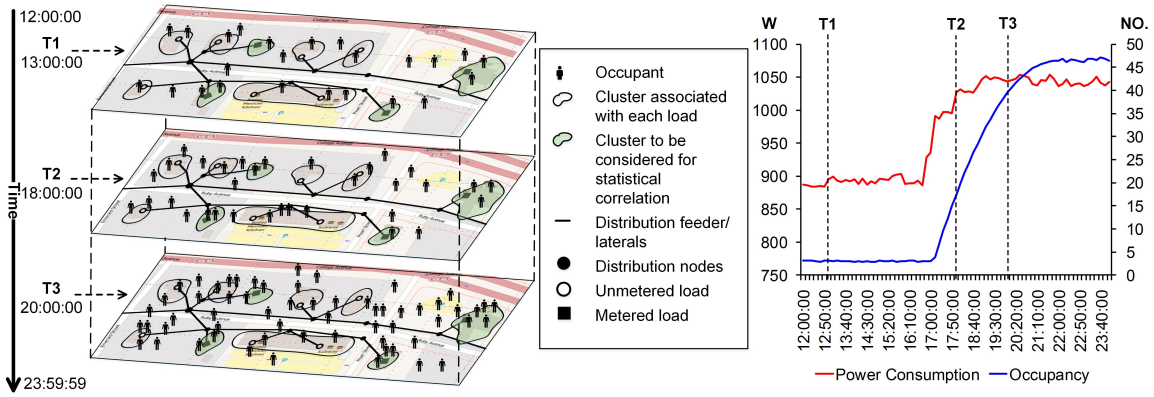

Figure 4.1: Ideal correlation of human movements and electricity consumption within a partial distribution feeder.

4.2 Enhanced Load Modeling Framework

Fig. 4.2 is the flowchart that presents the framework for this study. Variables considered in the formulation of the initial statistical regression model are the historical power consumption data and the statistics of occupancy. The input dataset can be obtained from the archived data of the previous period (daily, weekly, monthly, or seasonally) that is not influenced by weather change distinctly throughout a year. We select five basic statistical models as the candidates to establish the regression model. Once none of the models can meet the requirements of the validation, the proposed heuristic regression algorithm is applied to construct the initial combined regression model α. In order to reduce the estimation error associated with loading conditions, the weights are updated based on the structural deviation β [118].

The initial statistical model formulation is derived from $N_{o,a}$ and $P_{o,a}$, which is formulated as:

$$\alpha = \widehat{F}_t(N_{o,a}, P_{o,a}) \tag{4.1}$$

and the structural deviation of input of $N_{o,r}$ and $P_{o,r}$ is

$$\beta = \mu_{\Delta t}(N_{o,r}, P_{o,r}). \tag{4.2}$$

Therefore, the next snapshot of load information is evaluated and integrated into the calibrated model γ as follow:

$$\gamma = F_{t+\Delta t}(\alpha, \beta, \varepsilon_{\Delta t}). \tag{4.3}$$

This achieves incremental learning. The description of each module in Fig. 4.2 is

detailed in the following sections.

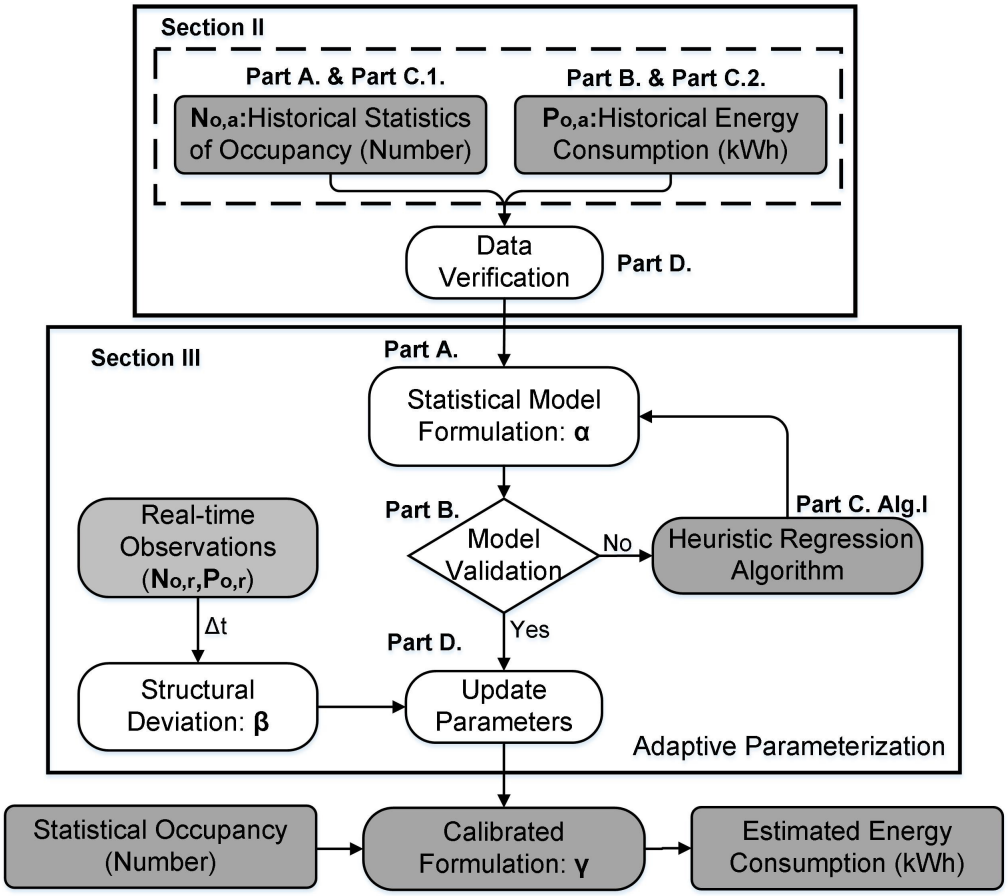

Figure 4.2: Flowchart of the proposed correlation framework.

4.2.1 Occupancy Data Availability

The daily statistical survey result of occupancy for a specific load with a regular energy usage schedule e.g., commercial, industrial, or agricultural loads, can be obtained based on the records of employee attendance sheets. This registration would be able to determine the duration that an occupant would consume the electrical energy.

All survey methods, however, can only provide a snapshot of information. Although motion sensors can detect users' motions near devices, massive system deployment can be cost-ineffective. Furthermore, the dynamic presence of occupants can be inferred by their cellular devices. To obtain the occupancy information based on this method, we make assumptions that: (i) every user in an observed load area holds only one unit and their devices are turned on, (ii) the estimated area is covered by Wi-Fi or cellular network, (iii) the assisted GPS (A-GPS) of mobile phones provides location accuracy up to 10 meters [119], and (iv) consumers' consent to share their location information.

4.2.2 Load Data Availability

The load survey is conducted based on consumption norm profiles that are observed within a time frame. The types of load profile are based on (i) individual household, (ii) consumer types, and (iii) the behavior of each load. Each profile can be collected throughout a year so that the unique individual profiles can be captured and established. Without IP-based metering devices, on-site readings of power consumption are undertaken by crews on a monthly basis. In addition, an image extraction method [83] can be utilized to obtain the real-time data from the eligible load as well. Also important is real-time information from supervisory control and data acquisition

(SCADA) [120]. This includes loading and voltage measurements by a limited number of metering devices installed along a distribution feeder. This feeder partition is useful in simplifying the regression analysis of system observability enhancement with a zonal approach.

Realistic scenario simulation may require a platform to integrate models for power systems, energy markets, building technologies, and the plethora of other resources that are becoming part of modern electricity production, delivery, and consumption systems. We implemented an agent-based approach by generating the load datasets using a multi-agent simulation engine including advanced algorithms that can determine the simultaneous state of millions of independent devices, each of which can be described by differential and algebraic equations [121]. The simulated outputs emulate data collection from the AMI that has been verified and validated in this study.

4.2.3 Agent-Based Modeling

An agent-based framework is proposed in this study for the sensitivity analysis because it enables a variety of third-party data management and utilization of other modules as part of the ecosystem under a simulation environment. The implementation of agent-based model utilizes the existing simulator [121] to support unbalanced

power flow for validation of sequential simulation test cases. The environment of open sources enables flexible enhancement in the simulator.

4.2.3.1 Occupancy Estimation

Suppose $A_{\text{Ave},o}$, which is the average area per occupant in an electrical load, only contains two elements '400' and '100,' which represents the square foot that one occupant needed in residential and non-residential areas, respectively. Then the average number of occupants in the target load generated from the agent-based model, O_a can be calculated by

$$O_a = \text{rnd}\left[\left(\frac{A_{\text{Total},o}}{A_{\text{Ave},o}}\right) \cdot N_{\text{floor}}\right]. \tag{4.4}$$

where $A_{\text{Total},o}$ is the total floor area of electrical load region and N_{floor} represents the number of floors in an electrical load.

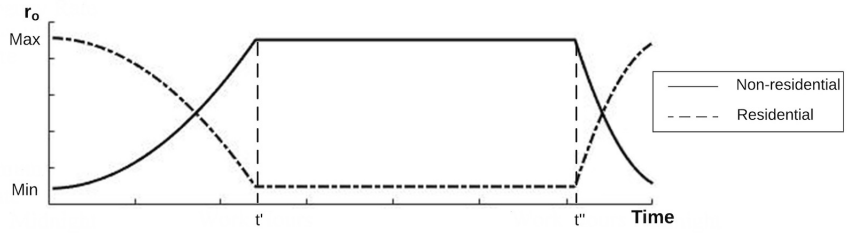

Figure 4.3: Ideal occupancy rate of a load area.

The aim of estimating O_a in a load is mainly to generate real-time occupancy data based on the dynamic occupancy rate. The solid curve in Fig. 4.3 demonstrates the ideal occupancy rate $r_o(t)$ in a non-residential load and the reverse of the curve displayed as the dashed line in the figure is the rate in a residential load. The values of the occupancy rate can vary according to the settings of maximum and minimum values and the different start t' and end t'' times of work hours. Overall, the increase and decrease of this rate are on the basis of quadratic function. The steady-state segment within this curve is not strictly linear. The O_t of a load is represented as (4.5):

$$O_t = O_a \cdot r_o(t). \tag{4.5}$$

4.2.3.2 Generation of Electrical Usage Datasets

Since the existing "`occupantload`" module in the simulator only consider the power consumption of heat gains [122], seven modules: 1) base load, 2) occupantload, 3) lights, 4) waterheater, 5) HVAC, 6) plugs, and 7) error consumption, are modified and designed in the proposed system [87]. A new property named as "`occupancy`" is added in modules: occupantload, lights, and plugs based on Eqs. (4.4) and (4.5).

The working schedule of each module is set according to human movement and behavior habits. Some default values such as the floor area in lights module and heat gains in "`occupantload`" are adjusted to meet simulation requirements. Consider a few occupants may neglect to turn off their appliances such as lights and laptops, or leave devices for charging. There may be extra power consumptions without any occupants' activities. The extra power consumption of each occupant is assigned with a number based on the statistics that may represent error consumption. The time frame to consider the error consumption should exclude the duration that the corresponding area is occupied. Therefore, simulated P_t from the model can be calculated by summing up power consumption generated from all modules and then minus the value from the error module.

A primary verification of the modified model is performed before the application. A comparison of the generated power consumption results between the initial simulator model and the modified model incorporated with the new property "Occupancy" is utilized to prove the validation of the changes in the source code. Once the difference of the time-varying consumption between these two is only influenced by the different number of occupants in a test case, the modified model can be applied for the simulation.

4.2.4 Data Verification

This module aims at assuring accuracy and integrity of input data for the initial statistical model α. This involves eliminating inconsistent or bad data points from the empirical data distribution. Using the proposed methods, a constant input of datasets is established under defined intervals. By setting this criterion, it is to compare the magnitude of the data input between its former and later ones. If the absolute values of these two differences are greater than the preset default error, the input data is considered as abnormal and will be ignored in constructing models. Some incomplete/missing or duplicated data points are verified in this integrity check module. This assures the consistencies of individual load consumption corresponding to occupancy information.

4.3 Regression Models Incorporating Estimated Occupant and Consumption Datasets

Human movement and its patterns are inter-correlated with respect to their existence in a specific load area as well as their contribution to power consumption. The sensitivity of the patterns to the number of occupants is studied within a timeframe

and how the hybrid regression model can be built by identifying the quantitative interdependency between these two. This section is a heuristic regression approach to establish a load profile based on the simulation of electrical loads and occupancy datasets. This is an iterative process that adapts the parameterization based on the observation.

4.3.1 Statistical Model Formulation

This module is the first part of the adaptive parameterization where it begins to determine and establish a statistical distribution module between the two data sources. A power and occupancy (PO) curve conversion is applied in this step. Under this representation, the occupancy information is shown in x-axis and power consumption corresponds to y-axis. The data points shown in diagram are the statistical distributions. An observation may have more than one cluster. For example, in Fig 4.4, the scatter diagram has three clusters with different time durations. The observation in the work hours and inactive time display a steady occupancy relatively correlating to power consumptions, while the other cluster deviates from the other two. Hence, a threshold is set to decide if the dataset needs to be partitioned according to the standard deviation of clusters. Ideally, the linear model should be the most direct approach to correlate the relationship between the variables. However, most actual data correlation may not exhibit linear dependencies, as a result, adaptive measures

with hybrid functions are necessary to capture the patterns precisely.

Determination of the regression coefficient can be estimated based on least squares analysis [123]. The unobserved random error of residual terms can be eliminated because the regression analysis conforms to statistical normal distribution of these error terms [123]. The identification of the load modeling of different data can be based on the comparison of fitting coefficients between the aforementioned five models.

4.3.2 Model Validation

The module follows from the previous section is to validate statistical parameterization from the preliminary formulation. There are four tests to be evaluated which are based on: (i) *R-squared test*, (ii) *F-test*, (iii) *significance test*, and (iv) *Geoffrey E. Havers statistic (GEH) test* [124].

R-squared test remains one of the important indicator in regression analysis is not determined entirely based on the fittest of datasets, but it also can be used to intereprete the variation and its dependent variables [88]. A general case with R-squared test of 0.75 would exhibit reasonable correlations [89].

The GEH method is another test to validate the collected data points and to determine the fitting outcome between the power consumption and the number of occupants in

a load area. The formulation of GEH is $G = \sqrt{\frac{2(m-c)^2}{m+c}}$. The benefit of using GEH is that the model provides an iterative method to constantly update the estimated value based on their statistical distribution in which the empirical norm output of GEH should be within a reasonable range, i.e., less than 5.0.

The hybrid model of statistical curve fitting would require to pass through the aforementioned four tests in order to form a reasonably justified hybrid model. The R-squared test must be executed after the F-test, the significant test, and the GEH test because these three tests are used to prove the input variables are able to construct the regression model or the model is meaningful. If any of the three fails, then the R-squared will not be initiated. Different test model would have the minimum threshold value, e.g., the R-squared test with at least 0.75. However, we use at least 0.8 value in our study as the threshold of R-squared test because it improves the fitness of the proposed model.

4.3.3 Heuristic Regression Algorithm

The heuristic regression algorithm proposed here is based on the finite mixtures of regression models for the best curve fitting. The Algorithm 1 is the pseudocode that shows the iterative procedure how the regression parameter adjustment is being adapted for each iteration. The heuristic regression algorithm is summarized as

follows:

1. Input the updated O_t and P_t;
2. Construct S by the OP paired points conversion that using O_t as the abscissa and P_t as the ordinate;
3. Withdraw S' randomly from S and the remaining data as R, where $S = S' + R$;
4. Apply each candidate model in CM_m for R. If any model in CM_m satisfy the criterion ν, return this candidate model as the result model;
5. If the given models cannot satisfy ν, withdraw the last λ (or percent) points to form R_{Part1} and R_{Part2};
6. Repeat $f_h(\cdot)$ for all new training sets to find the best candidate model in CM_m to satisfy the criterion ν of each set and combine all results as $\widetilde{M}^r_{R,t}$;
7. If the given models still can not satisfy ν for the new training set, repeat $f_h(\cdot)$ until M_h is constructed;
8. If there is not M_h till the last iteration, reduce the threshold value of the set partition requirement by ρ and repeat $f_h(R, CM_m, \nu)$;
9. Construct $\widetilde{M}^r_{R,t}$ with $c_b(\cdot)$;
10. Utilize cross-validated S' to perform each iteration. If the average error rate is reasonable, return $\widetilde{M}^r_{R,t}$; if not, repeat $f_h(R, CM_m, \nu)$ for S to calibrate the

model.

The threshold of the validation criterion will affect the number of iterations. The smaller the threshold value, the lesser the iterations it can be. The heuristic regression algorithm builds models based on the following:

$$\widetilde{M}^r(x,t) = \sum_{b=1}^{q} c_b(x) \cdot \widetilde{M}_b^r(x,t), \quad c_b(x) = \begin{cases} 1 & \text{if } x \in R_b, \\ 0 & \text{if } x \notin R_b \end{cases} \tag{4.6}$$

where $c_b(\cdot)$ is the input membership function in a subset while b is the index of a sub-model and q is the number of sub models in the heuristic regression model.

The model types of CM_m in different submodels could be the same but with different adaptive parameters. The cross-validation applied here is to avoid overfitting. Typically, the test set selects around 10% or 20% points of the dataset randomly for each iteration. The range of inference is based on the result from the heuristic regression model. The I_r is the power consumption inference range and e' is the average error of cross-validation that is updated iteratively and is in the range of $\widetilde{M}^r \pm e' \cdot \widetilde{M}^r$.

4.3.4 Update Parameters

In order to reduce the additional estimation error, the model should keep updating based on the real-time income observation. Under normal circumstances, a subtle change on the initial weights will be updated. However, under special circumstance such as a special event with a large number of occupants but normal power consumption is held in next time frame, the structure of the initial model could be changed. Therefore, constant update on the model based on real-time sampling input is necessary.

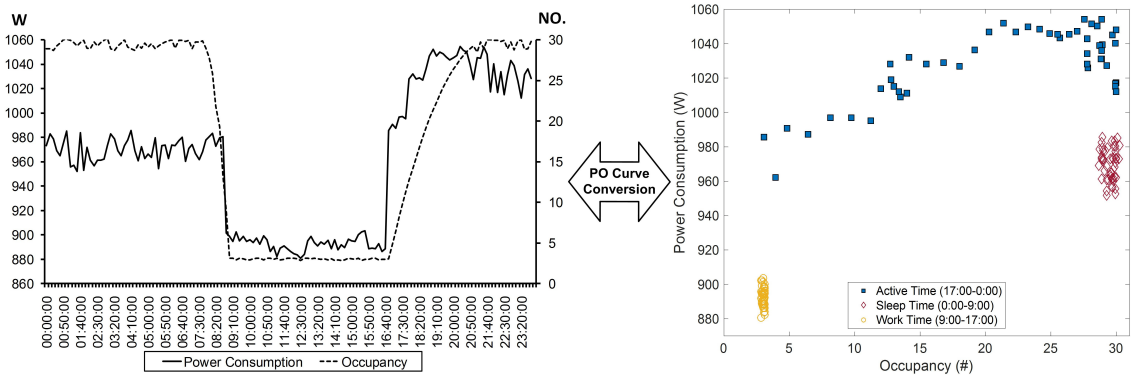

Figure 4.4: IEEE 13-node test feeder example of occupancy and residential load with OP pairs conversion.

4.4 Stimulated Case Study

This section is to validate the proposed heuristic model within the IEEE 13-node test feeder and a distribution network that covers a geographical region. The information is estimated for a day and the heuristic model is applied during the observation. The evaluations of simulated results and allocation factor of unmetered values based on occupancy is discussed in this section.

4.4.1 IEEE 13-Node Test Feeder

4.4.1.1 Test Case Description

An example of the time-series occupancy and power consumption in a residential area was illustrated in this part. The simulation topology is the IEEE 13-node test feeder and the test load area is the subfeeder of node 692 within the system as shown in Fig. 4.5. There are 10 residential loads connected with node 692 through a triplex meter to record relevant consumption information of this load area. The triplex meter provides a voltage to the house panel and the house provides a description of the current load to the triplex meter.

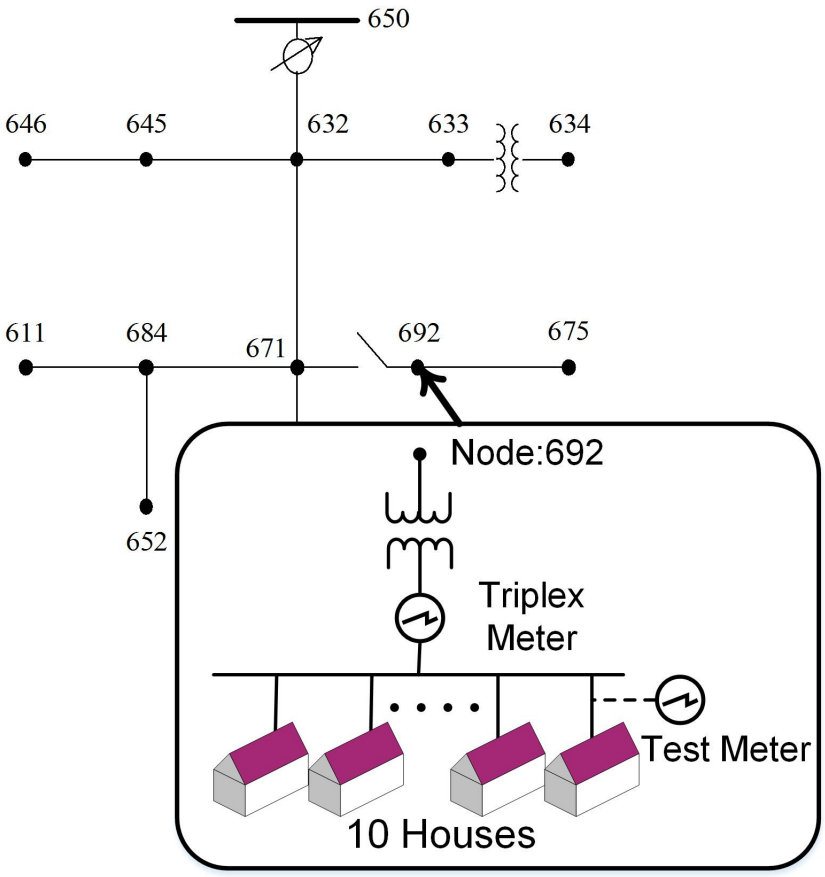

Figure 4.5: The detailed schematic diagram of the test subfeeder in the IEEE 13-node test feeder.

The module "`triplex_meter`" in this simulator provides a connection point between the residential and relative powerflow modules. Therefore, every house needs a triplex meter to connect to the electrical system (although multiple houses can connect to a single triplex meter). Conceptually, the triplex meter can be modeled as the customer-site meter, either electronically metered or manually obtained at sites. The secondary triplex system in this simulator only represents the circuit as seen in a residential house panel, or two 120-V circuits and one 240-V circuit. Commonly, the system is designed as the "`triplex_line`" connected to a "`triplex_node`" at the secondary

side of a center-tap or split-phase transformer [125].

The script consults the open source code of IEEE 13-node test feeder from [126]. We keep the default settings of other nodes and update the properties of node 692. The process of generating the electrical usage dataset in the agent-based model can be summarized as follows:

1. Initial the average area of loads in the subfeeder of node 692 at around 1,200 ft^2 with 30 occupants as the fixed households using Eq.(4.4).

2. The occupancy rate follows the curve trend shown in Fig. 4.3 and set the $t' =$ 9:00AM and $t'' =$ 5:00PM.

3. Add the new parameter "`occupancy`" to represent the occupancy rate (generated in Step 2) in modules: `occupantload`, `lights`, and `plugs`.

4. Set default values of power consumption for each occupant in modules `occupantload`, `lights`, and `plugs`.

5. The heating schedule is set with five time durations: 0:00-6:59, 7:00-8:59, 9:00-16:59, 17:00-21:59, and 22:00-23:59. The heating set points are 65°F, 70°F, 60°F, 65°F, 70°F, respectively. The cooling set point is set as 81°F.

6. The water demand schedule is set according to the occupancy rate and also with five time durations same as the heating schedule in Step 4. The water

demands are 0.01 gallons × "`occupancy`", 0.3 gallons × "occupancy, random between (0.1, 0.2) gallons × "occupancy, 0.3 gallons × "`occupancy`", and 0.5 gallons × "`occupancy`", respectively.

7. Set the heating schedule for HVAC (generated in Step 4) and water demand schedule for `waterhearter` module (generated in Step 5).

8. Generate the base load randomly between the range from 1.5VA to 2.5VA per half an hour and generate random error consumption for the test load area between 8VA and 10VA per hour.

9. Input the topology of the whole test feeder into the agent-based model.

10. Generate the time-series power consumption of the tested subfeeder every 10 minutes.

4.4.1.2 Establishing Statistical Models

The curve figure before the OP pairs conversion in Fig. 4.4 demonstrates the time-series occupancy and power consumption results in the subfeeder of IEEE 13-node test feeder using the proposed agent-based model for one day. The curve of the power consumption is obtained from the triplex meter connected with node 692. The left axis represents the power consumption and the right axis is the number of occupants.

The scatter diagram in Fig. 4.4 shows the correlation between the occupancy and the power consumption in the test area after data verification. The clusters of the three time durations have distinct distributions. The occupancy and power consumption during the work time (9:00-17:00) and inactive time (0:00-9:00) are relatively stable while the change of power consumption corresponding to the varying number of occupants is not distinct during these inactive time. However, the resultant trends of the active times (17:00-24:00) conform to the occupancy-power presumption that existence of additional occupants will result in more power consumption.

Five initial regression models are applied for the active time distributions as shown in Fig. 4.6 to construct the initial statistical model formulation in this case. Tables 4.1 is the summary of the initial validation without the cross validation. For all cadidate models under R-squared (column 1) in Table 4.1, none of them are larger than the required value 0.8 despite all of the other three tests meet their requirements. Therefore, the heuristic analysis is applied for curve-fitting adjustments.

Table 4.1
Sample data of initial validation summary of the test load area in IEEE 13-node test feeder.

Model	R-squared	F	Sig.	GEH
Linear	0.524	52.927	0	1.762
Logarithmic	0.630	81.767	0	2.309
Quadratic	0.731	63.736	0	1.323
Cubic	0.785	59.458	0	1.158
Exponential	0.526	53.369	0	1.614

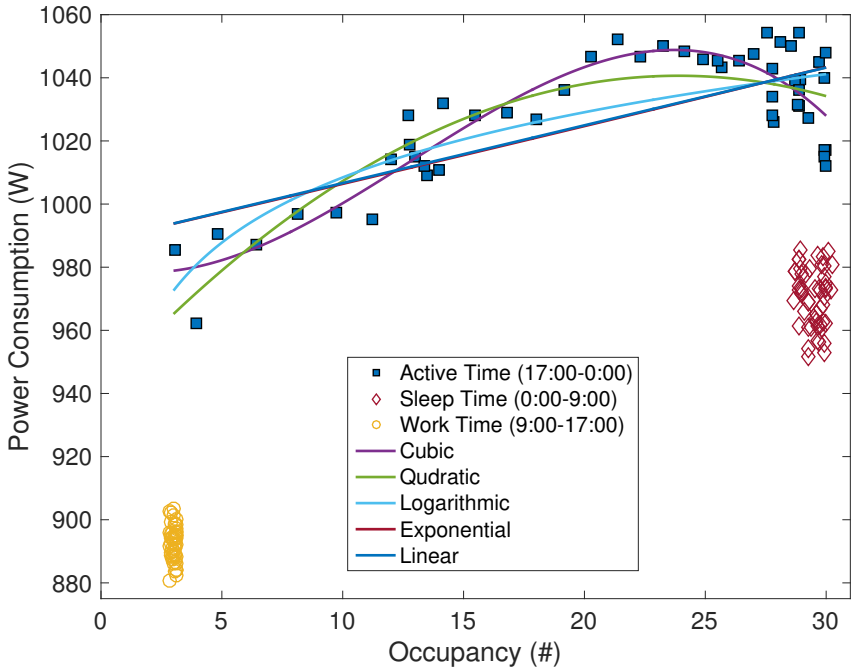

Figure 4.6: Initial regression analysis of one day sample data on test load area of IEEE 13-node test feeder.

4.4.1.3 Heuristic Regression Model

This section discusses parameter modification for the five regression models in accordance to the four satisfactory tests. According to Algorithm 1, the heuristic regression analysis is as follows:

1. Withdraw 10 points randomly from the points set as the test set for cross validation;

2. Since there is no candidate model that can satisfy the criterion, the resultant model will be a hybrid model;

3. Withdraw the last 10 percent points of the remaining points set to form two new training sets;

4. Apply the five candidate regression models for these two new training sets. These sets are used to find the best one to satisfy the criterion of each set. Then, combine all results into the hybrid model;

5. The hybrid model with two submodels (cubic and linear) is constructed;

6. Utilize the test set to perform the cross validation, since the process is reasonable, return the hybrid model as the initial statistical model.

The heuristic regression model is illustrated in Fig. 4.7. The hybrid model consists of two submodels: cubic and linear. The R-square value of each submodel are higher than 0.8 and the other three criterions are valid. Table 4.2 demonstrates the validation summary of the hybrid model for data distributions in the test load area. The cross validation column shows the maximum error rate of each model. Because the goodness of fitting we selected is 0.8, it can be observed that the error rate is relatively high. However, if the error rate is closer to 0%, it might be over-fitting and will not be useful for inference.

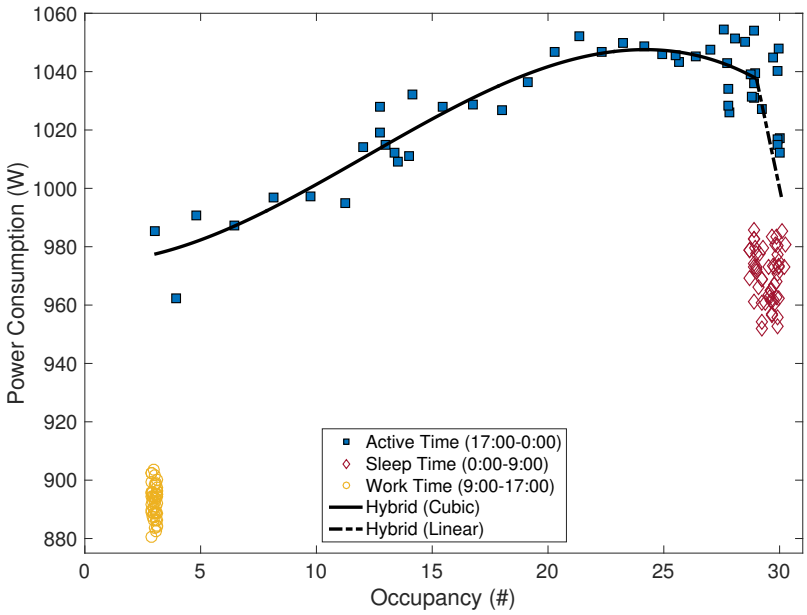

Figure 4.7: Heuristic regression analysis of one day sample data in the load area of IEEE 13-node test feeder.

Table 4.2
Validation summary of heuristic regression for the active time distributions in the load area of IEEE 13-node test feeder.

Model	F	Sig.	GEH	R^2	Cross Validation
Sub1	87.647	0	1.028	0.874	5.477%
Sub2	106.043	0	1.361	0.891	

⋆ Subs 1 & 2 represent 2 sub-models constructed by the heuristic algorithm.

4.4.1.4 Update Parameters and Study Results

In this case, we prefer to estimate the power consumption in 20:00 next day. In order to update the parameters in $\widetilde{M}^r_{(i,t)}$, real-time data of active time is generated from 17:00 to 19:50 in the next day as $\widetilde{M}^r_{(j,\Delta t)}$ to the initial data set. As illustrated in Fig. 4.8, the structure of the hybrid model remain unchanged while the parameters are

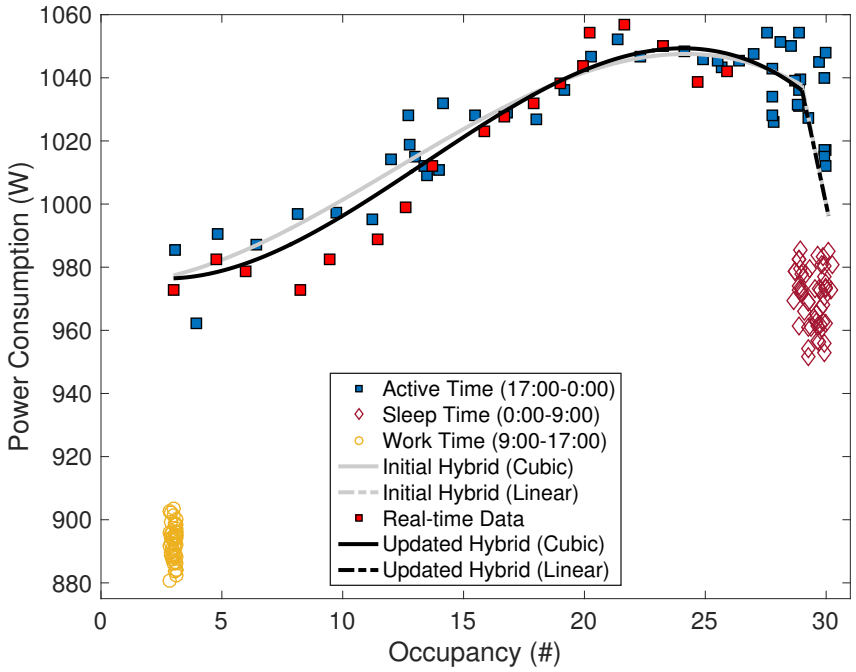

Figure 4.8: Updated regression model of upper branch in the test load area of IEEE 13-node test feeder.

updated following the real-time data input.

A test meter is connected with a load randomly in this subfeeder. Since the consumption relative parameters of each load are set as the same in the agent-based simulator, and there are 10 loads in this area, the power consumption percentage of each residence is around 10% in the subfeeder. Table 4.3 shows the error rate analysis between the heuristic model and the value generated from the following time index. The model selects the time as 20:00 with 25 occupants. The rows with "Generated" and "Hybrid" correspond the estimation error of the whole subfeeder and the error rate of the heuristic regression model is 0.535%, which is obviously in an acceptable range. The other two rows demonstrate the error rate analysis between the estimation

and the value gathered from the test meter in the random residential load. The error rate is 1.72%.

Table 4.3
Error rate analysis of the IEEE-13 nodes test feeder load in a concentrated interval.

Name	Occupancy (#)	Power Consumption (W)	Error Rate (%)
Generated	25	1,043.34	-
Hybrid	25	1,048.92	0.535
Metered	-	106.73	-
Estimated	-	104.89	1.72

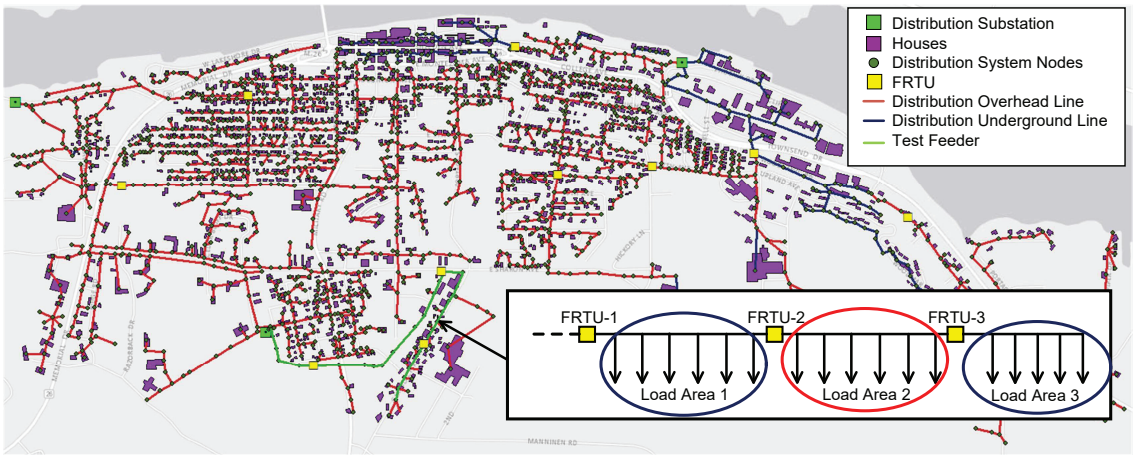

Figure 4.9: The test distribution system in geographical proximity of two focused areas.

4.4.2 An Electrical Distribution System

4.4.2.1 Test Case Description

Fig. 4.9 is the geographical region of the distribution test network. There are 3 substations, 1 665 distribution nodes, and 1 644 houses (include residential and non-residential). The conductor data for the overhead line are selected by the types of 556 500 26/7 ACSR and 4/0 6/1 ACSR while the configuration of conductor for the underground line is 250 000 AA. The detailed schematic diagram in Fig. 4.9 illustrates a test feeder with three feeder remote terminal units (FRTUs) in this network and the loads are indicated by arrows. The heuristic regression model is tested in Load Area 2 in this feeder. This area is selected because all of the loads in this area are non-residential while the load types in Load Areas 1 and 3 are hybrid. There are six subfeeders connected to this load area. The electricity usage of each subfeeder is equal and then the consumption percentage of each subfeeder is around 16.67%. In relative terms, the residential load is more sensitive than the non-residential load because an occupancy change can be observed with apparent variation in their consumption.

The process of generating the power consumption datasets is similar as the IEEE 13-node test feeder example. The settings of t', t'', heating and cooling set points, and water demand schedule are same as the previous case. However, we initial the

average areas of all loads in the system according to the geographical information gathered from the related software [127]. The average area of the Load Area 2 is around 24 000 ft^2 while there are 240 occupants are fixed personnel based on Eq. (4.4). In addition, the base load changes from 4.5VA to 5.5VA per half an hour and the error consumption is generated randomly from 5VA to 8VA per half an hour. Since the temperature may affect the simulation result, this study is assumed to be in a season that does not fluctuate a significant change in temperature.

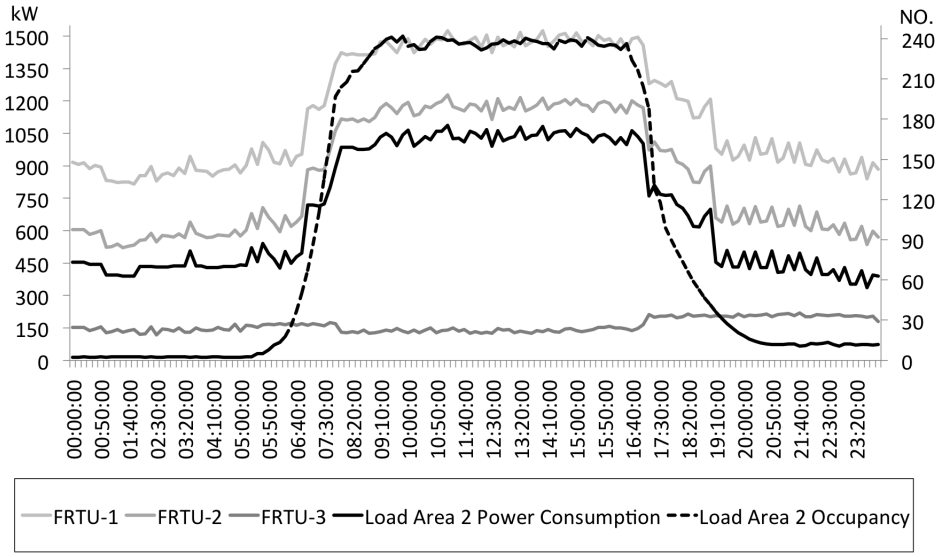

Figure 4.10: Occupancy and power consumption results in the non-residential load area (Load Area 2).

4.4.2.2 Establishing Statistical Models

The time-series power consumption results in these three FRTUs are shown in Fig. 4.10. Therefore, we can determine the consumption difference in the non-residential

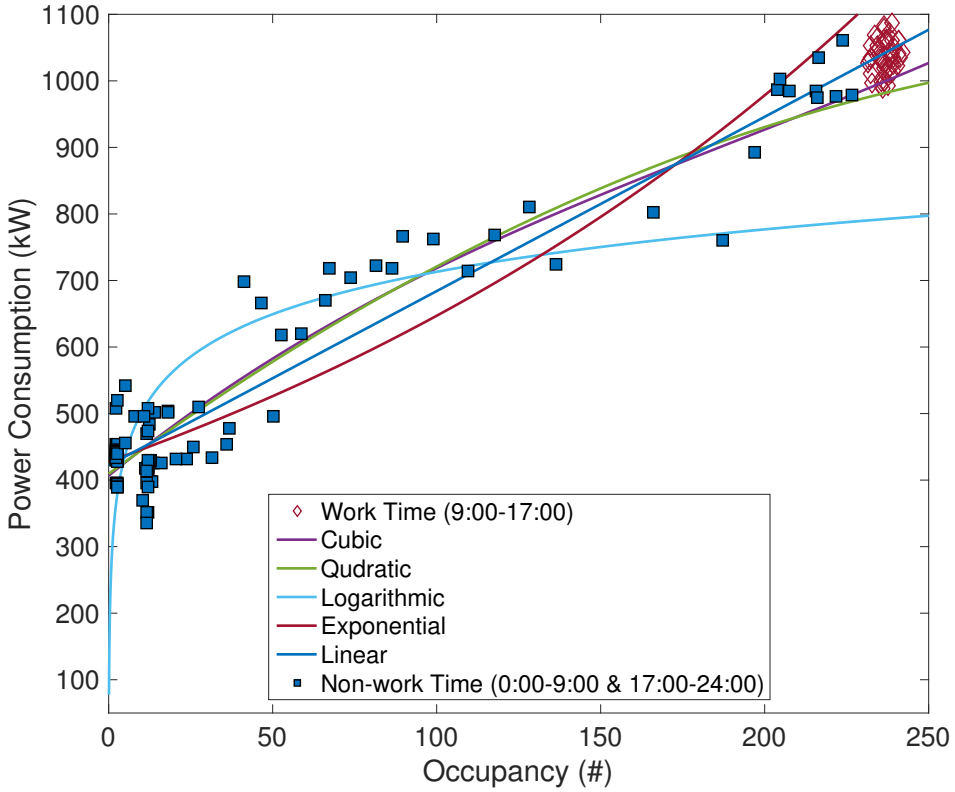

Figure 4.11: Initial regression models of one day sample data in the non-residential load area (Load Area 2).

area (Load Area 2) by subtracting the FRTU-3's values from FRTU-2. The result and the occupancy curve are also shown in Fig. 4.10.

The OP paires in Fig. 4.11 demonstrates the sample data for the Load Area 2. Five candidate regression models are applied in this study based on Section III.A. To improve the precision of regression models, the observed data cluster represents the correlation during work hours with relatively stable power consumption is not analyzed. The simulation on this load area is based on the points distributions in non-work hours.

Tables 4.4 is the summary of initial validation without cross validation. Each column shows the corresponding parameters to the five regression models. For all regression models are less than 0.8 shown in Tables 4.4 under R-squared (column 1). This exhibits reasonable results for the R-squared test. However, for the other three tests, each of the columns shows consistently that they are more within their threshold values; column 2 satisfies the threshold from F-distribution table, column 3 is with less than 0.05, and column 4 is less than 5. Under this circumstance, the heuristic algorithm continues to adjust until it reaches a convergence.

4.4.2.3 Heuristic Regression Model

According to Algorithm 1, we have withdrawn 10 points randomly as Γ and the algorithm will withdraw the last 20 percent (λ) points to process the heuristic regression algorithm for training generated sets, respectively. The algorithm repeats $F_h(\cdot)$ until it has met the criteria. The cross validation is performed to check e_{cv} of the constructed hybrid model. If the result is reasonably established, the hybrid model will be utilized for inferring power consumption in the following time stamp. As the database of R grows with real-time observation input, the regression model is updated constantly to refine the regression model for prediction.

Figs. 4.12 illustrates the heuristical regression model for the non-residential area (Load Area 2). The hybrid regression model consists of two subsets. The R-squared

Table 4.4
Sample data of initial validation summary in the non-residential load area (Load Area 2).

Model	R-squared	F	Sig.	GEH
Linear	0.715	799.879	0	2.11
Logarithmic	0.659	182.050	0	1.145
Quadratic	0.781	423.345	0	1.091
Cubic	0.783	284.988	0	1.104
Exponential	0.678	673.629	0	2.120

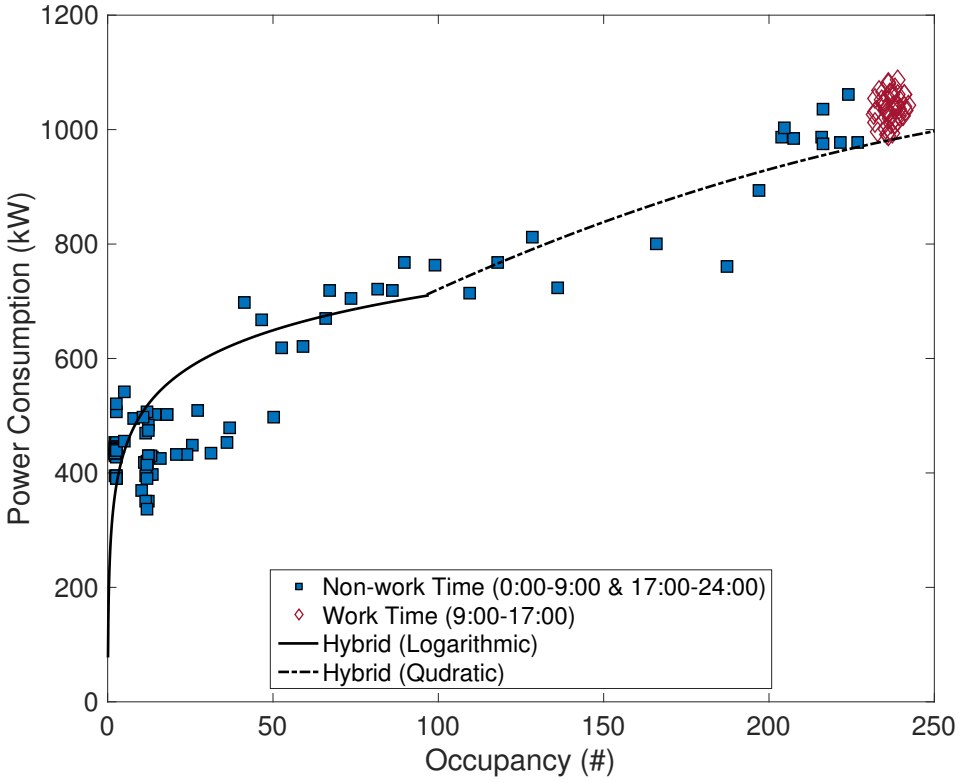

Figure 4.12: Heuristic regression analysis of one day sample data on the residential load area.

value of each submodel are higher than 0.8 and the other three criterions are valid. Table 4.5 shows the validation summary of heuristic regression model for the the residential load area. The model types of submodels are logarithmic and quadratic.

Table 4.5
Validation summary of heuristic regression for the non-residential load area (Load Area 2).

Model	F	Sig.	GEH	R^2	Cross Validation
Sub1	803.357	0	2.119	0.895	5.468%
Sub2	465.821	0	2.061	0.802	

$\star$ Subs 1 & 2 represent the 2 sub-models constructed by the heuristic algorithm.

4.4.2.4 Study Results

Similar to the previous case, we connected a test meter with a subfeeder randomly in this test load area. The regression model is setup with parameters of time as 7:40 and 180 occupants. The structure of the regression model remains unchanged while the parameters of models are updated following the real-time data input. Table 4.6 illustrates the error rate analysis between the heuristic models, and the value captured from the following time index by the test meter.

The error rates of heuristic regression models are under 5%, which are in an acceptable range based on (4.7).

$$I_r \in \left(\widetilde{M}^r \pm e_{cv} \cdot \widetilde{M}^r\right). \tag{4.7}$$

where $M_h(x)$ is the result from (4.6) and e_{cv} is the maximum error rate of the cross

Table 4.6
Error rate analysis of the test load in concentrated interval.

Name	Occupancy (#)	Power Consumption (kW)	Error Rate (%)
Load Area 2	180	791.104	-
Hybrid	180	801.436	1.306
Metered	-	136.393	-
Estimated	-	133.572	2.068

validation.

The error includes the statistical error in data estimation and the error might be caused by other factors such as the existing internal modules and default parameters in the simulator for estimating power consumption. In practice, the the percentage of energy consumption for each subfeeder in a load area can be estimated based on archived data.

Algorithm 1 Heuristic Regression Algorithm

Require:

S: A set of OP paired points based on O_t and P_t.

S': A test set for cross-validation.

R: A training set to construct a regression model.

$f_h(\cdot)$: Iteration function of heuristic regression.

CM_m: A set of candidate models, where m is the index for candidate models, $m = 1, 2, \cdots$

ν: Test criteria to validate a regression model.

n: Number of elements in R.

$n_{\min}$: Minimum number of elements in R can partition set in the heuristic regression algorithm.

λ, ρ: Adjustment parameters.

Input: O_t, P_t

Construct S.

Iteration Process: Withdraw S' from S and left R.

if $f_h(R, CM_m, \nu)$ satisfy ν **then**

  $\widetilde{M}^r(R,t) \leftarrow f_h(R, CM_m, \nu)$;

  return $\widetilde{M}^r(R,t)$;

else

  while $n \geq n_{min}$ **do**

    $R_{\text{Part1}} \leftarrow R(1, 2, \cdots, \lambda)$,
    $R_{\text{Part2}} \leftarrow R(\lambda + 1, \cdots, n)$.

    Do $f_h(R_{\text{Part1}}, CM_m, \nu)$, $f_h(R_{\text{Part2}}, CM_m, \nu)$

$$\widetilde{M}^r(R,t) = \begin{cases} \widetilde{M}^r_{\text{Part1},t} = f_h(R_{\text{Part1}}, CM_m, \nu), R \in R_{\text{Part1}} \\ \widetilde{M}^r_{\text{Part2},t} = f_h(R_{\text{Part2}}, CM_m, \nu), R \in R_{\text{Part2}}. \end{cases}$$

  end while

end if

if No $\widetilde{M}^r(R,t)$ generated. **then**

  Reduce $n_{\min}$ by ρ;

  Do $f_h(R, CM_m, \nu)$

end if

Cross-Validation: $\widetilde{M}^r(R,t)$

Using S' to evaluate cross-validation.

if Average cross-validation is reasonable **then**

  return $\widetilde{M}^r(R,t)$.

else

  $R \leftarrow S$;

  $f_h(R, CM_m, \nu)$

end if

Chapter 5

Enhancement of Load Modeling by Correlating Between Occupancy and Consumption

5.1 Introduction

Although the rough requirement for accurate load models has been achieved by electrical researchers and engineers, more research is imperative to update existing load models and understand characteristics of modern loads with emerging smart grid technologies such as distributed generators (DGs), electric vehicles (EVs), and

demand-side management (DSM) [91]. The most uncertainty and difficulty of load modeling come from the lack of real-time measurements and detailed load information [48, 49, 81]. Due to load diversity and variation, an ongoing survey on specific profiles of unmetered loads has been conducted constantly [94, 95]. For load model structure development, more sophisticated models that balance flexibility and complexity are needed. Since load consumption is time-varying due to human behaviors, different load models may be found in different time periods. Conventional load modeling methods using measurement data in a certain period may not be able to capture time-varying load behaviors and lack generalizability. More research is needed to develop advanced algorithms to perform online load modeling using the real-time data. After developing new load models, they should be integrated in power system analysis.

Since some utilities do not implement full smart meter deployment, approximate methods have been used to gauge the weighting factors of individual loads which can be studied from a survey or to estimate using allocation factors based on transformer ratings associated with a root node of the metering point [95] are utilized for planning purpose, while the operational load models are typically constructed with a combination of the geographic information system (GIS) and the data management system (DMS) information. Hence, the operational load models for each household are often updated manually by their crew. Such labor-intensive statistics may not describe their consumption activities.

Prediction and utilization of energy usage at appliance level can be estimated and improved [108, 109, 110, 111, 112]. Control management of local electricity usages is further investigated based on offline occupancy and demand response with improved methods [97, 101, 106, 113, 114, 115, 116]. The statistical correlation between occupants' movement and their electricity consumption associated with metering devices has been shown in last chapter. This chapter is prefer to apply the realistic load metering datasets to validate the proposed correlation model and method.

5.2 Enhanced Load Modeling Framework

Fig. 5.1 is the flowchart that presents the framework for this study. Variables considered in the formulation of the initial statistical regression model are the OP paired points generated from the historical power consumption data and the statistics of occupancy. The input dataset can be obtained from the archived data of the previous period (daily, weekly, monthly, or seasonally) that is not influenced by weather change distinctly throughout a year. Some selected basic statistical models (such as linear, qudratic, cubic, exponential, etc.) as the candidates to establish the regression model. Once none of the models can meet the requirements of the validation, the proposed heuristic regression algorithm is applied to construct a priorio estimation regression model, $\widetilde{M}^r_{(i,t)}$, for OP paired points i during time interval t, which represents the observation time interval, i.e. sampling time interval. In order to reduce

the estimation error associated with loading conditions, the weights of the regression model are updated based on the real-time data points j input within the time shift Δt [118]. The estimation model based on the archived data is as 5.1 and the calibrated estimation model is formulated as 5.2:

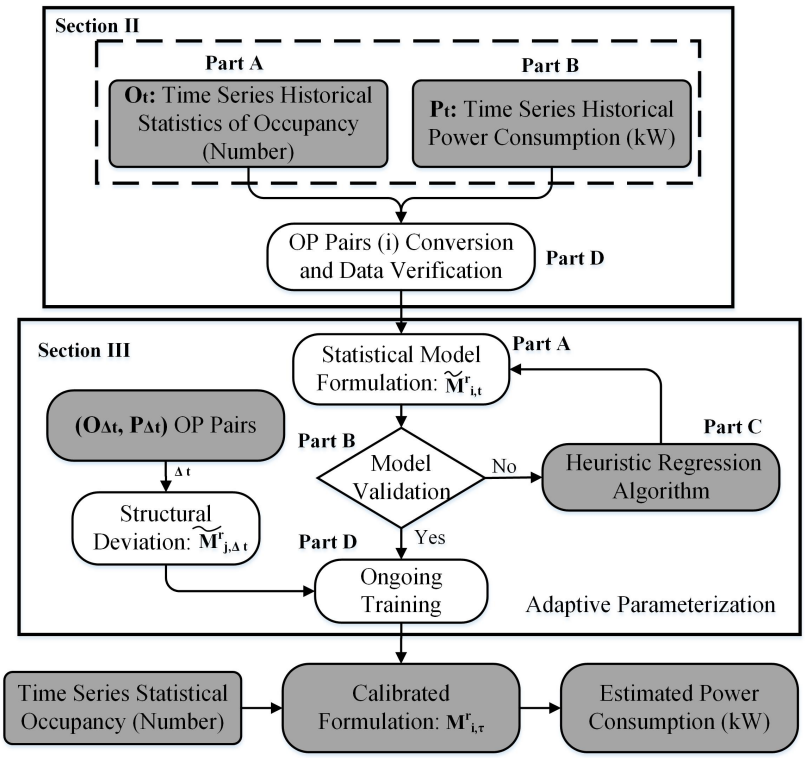

Figure 5.1: Flowchart of the proposed correlation framework.

$$\widetilde{M}^{r}_{(i,t)} = f[O(t), P(t)] + w_{(t)} \tag{5.1}$$

$$M^r_{(i,\tau)} = f[O(\tau), P(\tau)] + w_{(\tau)} = \widetilde{M}^r_{(i,t)} + \widetilde{M}^r_{(j,\Delta t)} \tag{5.2}$$

where τ represents the aggregated time intervals and $\tau = t + \Delta t$. $O(t)$ and $P(t)$ are data sets of archived statistics of occupancy and human related electrical loads during time interval t. w is random (white) disturbance (or process noise) vector which is assumed to follow a normal distribution with zero mean, that $w_{(t)} \sim \mathcal{N}(0, Q_{(t)})$ where $Q_{(t)}$ is the covariance. f is the system vector function 5.3 where n is the number of sub-models generated from the regression process and $f_k()$ is any nonlinear or linear function.

$$f = \begin{bmatrix} f_1() \\ f_2() \\ \vdots \\ f_k() \end{bmatrix}. \tag{5.3}$$

Once the establishment is made, it is used as the ongoing training to estimate the consumption for the current time index. Theoretically, only the a priori model $\widetilde{M}^r_{i,t}$, reflecting prior survey data and surveillance information up to the previous time

index, is available before performing the real-time estimation. Therefore, the real-time demand $M_{i,\tau}^{r}$ in the following study is modeled as a linear combination of the a priori estimate, the real-time estimation, and the random disturbance.

5.2.1 Occupancy Data Availability

As mentioned in Chapter 4, the study of human mobility can regularly infer traffic conditions from the cellular devices by tracking their positions [84, 85, 86, 96]. A regression framework can be established to enhance the dynamic behavior of occupants [97, 98, 99]. The dynamic presence of occupants can be inferred by their cellular devices.

5.2.2 Load Data Availability

Ideally, each consumer can be collected real-time data through smart meters so that unique individual profiles can be established. Without IP-based metering devices, on-site readings of power consumption are undertaken by crews on monthly basis. To improve the reading efficiency, an image extraction method [83] can be utilized to obtain the real-time data from the eligible load. Also important is real-time information from supervisory control and data acquisition (SCADA) [120]. This includes

loading and voltage measurements by a limited number of metering devices installed along a distribution feeder. This feeder partition is useful in simplifying the regression analysis of system observability enhancement with a zonal approach.

5.2.3 Assumptions

5.2.3.1 Assumptions of Data Collection

The effects of occupancy are sometimes not noticeable. There are numerous examples such as large industrial facilities or agricultural facilities and automation that can be insensitive to the existence of occupants at the site, e.g., server farms, refineries and chemical plants, cold-storage facilities, irrigation pumps. In addition, in some commercial and residential buildings, load differ may often be closely related to weather but the related adjustment is controlled by the smart thermostats. During the occupancy data collection, many consumers may not be willing to provide their locations or information. In view of above, some assumptions should be declared in this part:

† The focus of this research addresses residential or non-residential loads with time-varying consumptions.

† The sensitivity of occupant existence is subject to the type of loads. The electrical load types in this study are sensitive to consumers' existences within the

building or otherwise.

† There are no automatic adjustment device or smart equipment such as smart thermostats in the test area.

† There are several metering points within a feeder that would provide the data acquisition of the lumped load. The utilities have the metering data from SCADA/DMS system. These could include FRTU and substation RTU.

† The correlation approach is to harness the metering information together with the occupancy datasets. A data exchange between utility and data providers (aggregated data sources on occupancy) would establish this framework.

† If this paradigm were to be implemented, utilities would have to exchange information from the data providers that could constantly obtain the occupancy information from the provider. The frequency of data exchange could be every 10 minutes.

† Every consumer consents to share their location information and holds only one cellular device. Their devices are turned on. The estimated area is covered by Wi-Fi or cellular network and the assisted GPS (A-GPS) of mobile phones provides location accuracy up to 10 meters [119].

5.2.3.2 Assumptions of Regression

Most statistical tests rely upon certain assumptions about the variables used in the analysis. When these assumptions have not met the results may not be trustworthy. Several assumptions of regression are "robust" to the violation and others are fulfilled in the proper design of a study. Specifically, the assumptions of the relationship of variables, the reliability of measurement, and statistical characteristics should be considered in this section.

† The relationship between the independent and dependent variables is linear. The linearity assumption can best be tested with scatter plots. The 2D OP pairs scatter diagram is an example illustrated in Fig. 5.2.

† All variables to be multivariate normal in this analysis.

† Multicollinearity occurs when the independent variables are not independent from each other. An assumption is that there is little or no multicollinearity in the input variables.

† Autocorrelation occurs when the residuals are not independent from each other. Another assumption is that there is little or no autocorrelation in the input variables.

† The last assumption the regression analysis makes is homoscedasticity.

5.2.4 Data Verification

This module aims at assuring the accuracy and integrity of the input data for the priorio statistical regression model $\widetilde{M}^r_{(i,t)}$. This involves eliminating inconsistent or bad data points from the empirical data distribution. Using the proposed methods, a constant input of datasets is established under defined intervals. By setting this criterion, it is equivalent to a moving average filtering process by replacing each discrepant data point with the average of the neighboring data points defined within a preset span $(2N_{nei}+1)$:

$$y^s_{(i)} = \frac{1}{2N_{nei}+1} \cdot (y_{(i+N_{nei})} + y_{(i+N_{nei}-1)} + ... + y_{(i-N_{nei})}). \tag{5.4}$$

$$y_{(i)} = \begin{cases} y_{(i)}, & |y_{(i)} - y^s_{(i)}| \leq \theta, \\ y^s_{(i)}, & |y_{(i)} - y^s_{(i)}| > \theta. \end{cases} \tag{5.5}$$

where $y_{(i)}$ is the ith OP paired point and N_{nei} is the number of neighboring data points, i.e. OP paired points, on either side of $y_{(i)}$. $y^s_{(i)}$ represents the smoothed value from moving average filter for the ith OP paired point. θ is the preset default threshold for data verification.

The verification compares the magnitude of the data input and the smooth data generated from Eq. 5.4. If the absolute values of these two differences are greater than the preset default error, the input data is considered as abnormal and will be replaced by the smooth data as shown in Eq. 5.5 to avoid data lacking. Some incomplete/missing or duplicated data points are verified in this integrity check module. This assures the consistencies of the individual load consumption corresponding to the occupancy information.

5.3 Regression Models Incorporating Estimated Occupant and Consumption Datasets

Human movements and their patterns are inter-correlated with respect to their existence in a specific load area as well as their contributions to the power consumption. The sensitivity of the patterns to the number of occupants is studied within a timeframe and how the hybrid regression model can be built by identifying the quantitative interdependency between these two. This section is a heuristic regression approach to establish a load profile based on the simulation of electrical loads and occupancy datasets. This is an iterative process that adapts the parameterization based on the observation.

5.3.1 Statistical Model Formulation

The OP paired points conversion is applied in this section. Under this representation, the occupancy information is shown in the x-axis and the power consumption corresponds to the y-axis. A time series 3D observation may also be illustrated with time index as the z-axis. Fig. 5.2 shows OP paired points in 2D and 3D demonstration, separately. The figure displays gathered data from five workdays, every 10 minutes represent one time index in this study. Since the metering device in a single load occurred a few faults during the test period, some abnormal readings exist in the time series consumption change as an example indicated with an arrow, while the corresponding OP paired points observations also display inconsistent points (indicated with an arrow).

After the data verification process mentioned before, the load modeling identification of different data based on the comparison of fitting coefficients (least squares values) between the selected candidate regression models. Under normal circumstance, adaptive measures with hybrid functions generated from the heuristic regression algorithm rather than a linear model are necessary to capture the patterns precisely.

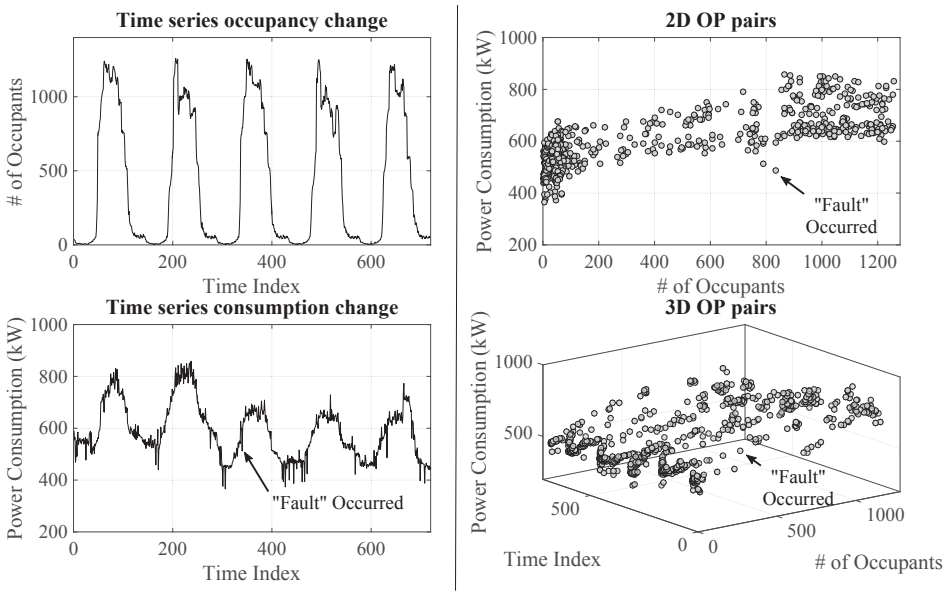

Figure 5.2: Time series occupancy and power consumption change and converted 2D and 3D OP paired points.

5.3.2 Model Validation

As same as in the previous Chapter, there are four tests to be evaluated which are based on (i) the *R-squared test*, (ii) the *F-test*, (iii) the *significance test*, and (iv) the *Geoffrey E. Havers statistic (GEH) test* [124]. The R-squared value is also set as 0.8.

5.3.3 Heuristic Regression Algorithm

The heuristic regression algorithm has been proposed in last Chapter as Algorithm 1 is the pseudocode that shows the iterative procedure how the regression parameter adjustment is being adapted for each iteration.

5.3.4 Ongoing Training

In order to reduce the additional estimation error, the model should keep updating based on the real-time income observations. Under normal circumstances, a subtle change in the initial weights will be updated. However, under special circumstance such as a special break with a relatively few number of occupants but the normal power consumption is held in the next time frame, the structure of the initial model could be changed. Therefore, the constant update on the model based on real-time sampling input is necessary.

5.4 Stimulated Case Study

This section is to validate the proposed correlation analysis with the realistic load metering datasets of an observed area. The occupancy information is estimated based on the registration states and the area occupied schedule from Fall 2017 semester datasets [128].

5.4.1 Test Case Description

This observed area has a substation and four backup generating units. There are three distribution feeders connected to the substation and each lumped load has a primary and backup connection to one of those feeders. Two lumped loads under one same feeder with five transformers are selected from the observed area to perform the correlation analysis. Since all of the circuits associated with these loads have IP-based energy meters and the lumped loads have high traffic that has the most human movements, that would be helpful for us to establish statistical study using the proposed framework. Furthermore, the lumped loads contain one of the highest consumption across the whole observed area.

5.4.2 Time Windows of Metering and Occupancy Datasets

The time windows considered in this data gathering are based upon the consistent intervals that are captured for occupancy and metering information. To begin with this process, we selectively picked a sample of datasets based on the weekly basis that would typically use for regressions analysis. Under same weather season, the weekly sample demonstrates the consistent patterns that can be built as a norm profile for the lumped loads with metering and occupancy data. For an example,

the comparison between past week with current one would always show consistent information that can be captured to build the occupant-metered profile relations. However, the monthly comparison may not be a reasonable data comparison for establishing a reference profile statistically. The time windows of Fall 2017 datasets are between October 2, 2017, and October 6, 2017. The number of enrollments reflected on this time period has shown relatively stable.

The default value of how the power consumption (kW) is captured is every 10 minutes. The connection between the distribution management system of the observed area and the devices are being transferred using IP-based Internet link module (ILM). The wired connection assures higher data reliability. The occupancy information is carefully mapped to the registration records from the campus database to ensure the classroom location and its association of energy consumption with a particular area. This estimate includes not only the students who were taking the classes, but also the number of faculty members, staffs, and graduate students who have their routine to these areas. This number varies loads to loads that can be insensitive to its others. Since weekdays and weekend may have statistical fluctuation, we only use five weekdays data in this study.

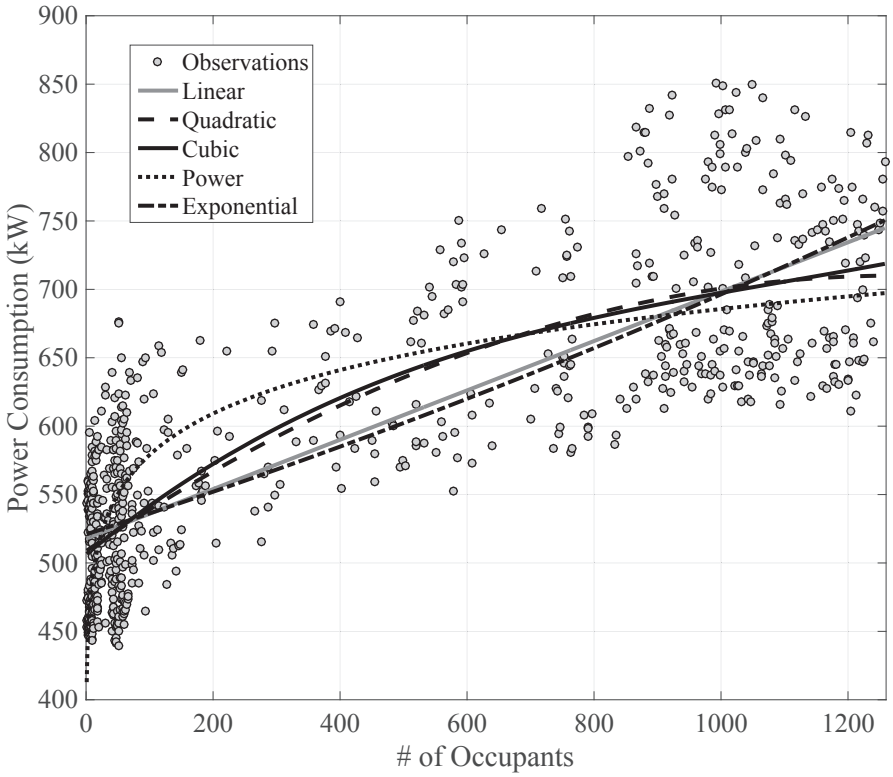

Figure 5.3: Initial regression analysis of 2D sample data on test lumped loads.

5.4.2.1 Establishing Statistical Models

Fig. 5.3 illustrates the filtered sample data of the lumped loads in the 2D presentation. The x-axis represents the number of occupants and the y-axis is the power consumption (kW). Five regression models: 1) linear, 2) quadratic, 3) cubic, 4) power, and 5) exponential, are selected in this case to construct the initial statistical model formulations. Tables I is the summary of the initial validation without the cross-validation. For all candidate models under R-squared (column 1) in Table 5.1, none of them are larger than the required value 0.8 despite all of the other three tests

meet their requirements. Therefore, the heuristic analysis is applied for curve-fitting adjustments.

Table 5.1
Sample data of initial validation summary of test lumped loads.

Model	R-squared	F	Sig.	GEH
Linear	0.661	52.927	0	1.762
Quadratic	0.678	63.736	0	1.323
Cubic	0.679	59.458	0	1.158
Power	0.641	81.767	0	2.309
Exponential	0.652	53.369	0	1.614

Fig. 5.4 (a) demonstrates the 3D regression analysis based on the sample data of the lumped loads. The x-axis represents the number of occupants, the y-axis is the time index, and the z-axis represents the power consumption (kW). Apply the cubic degree polynomial of two independent variables, x and y, to do the curve fitting, a statistical model formulation of a dependent variable can be constructed. The corresponding residuals plot (Fig. 5.4 (b)) shows the mean square values of residual points. Even though several points with relatively high mean square errors, the overall residual mean square error is around 32.5 in this study. Theoretically, the residual mean square error closer to 0 indicates a fit that is more useful for prediction, but it will cause overfitting. Therefore, we can define an estimation range based on the calculated value from the formulation and the residual mean square error. Since the R-squared value for this model is 0.8239 and all of the other three tests meet their requirements, the 3D statistical formulation can be utilized for the consumption estimation.

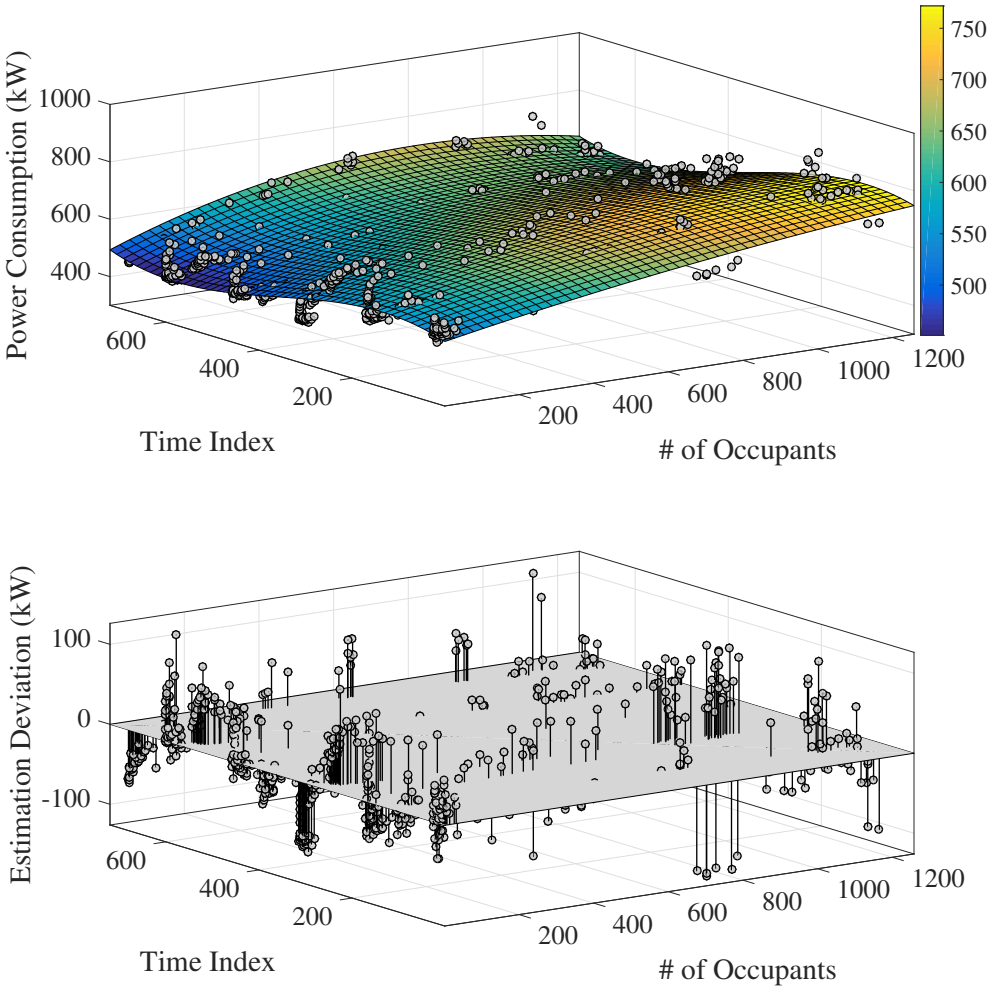

Figure 5.4: Regression analysis of 3D sample data on test lumped loads.

5.4.2.2 Heuristic Regression Model

This section discusses candidate model selection, combination, and parameter modification according to the four satisfactory tests. According to the Algorithm 1, the heuristic regression analysis is as follows:

1. S_{cv} is set as 10, which means withdraw 10 points randomly during each iteration from the points set as the test set for cross-validation;

2. Since there is no candidate model that can satisfy the criterion, the resultant model will be a hybrid model;

3. n_{min} is set as 10 and ρ is set as 1;

4. λ is set as 20, which means withdraw the last 20 percent of the remaining points set to form two new training sets;

5. Apply the five selected candidate regression models for these two new training sets;

6. Since there is still no candidate model that can satisfy the criterion during the first iteration, partition each set as two new sets according to λ and continue the iteration function;

7. Once the regression fitting of each subset satisfies the criterion, combine all results into a hybrid model;

8. During each iteration, utilize the test set to perform the cross-validation and calculate the average error rate. Since the value is reasonable, return the hybrid model as the heuristic statistical model.

The heuristic regression model is illustrated in Fig. 5.5. The hybrid model consists of 69 submodels and the types of this hybrid model contain linear, quadratic, and

cubic. The average R-square value of the hybrid model is 0.8132 and each submodel's R-square value is higher than 0.8. The other three criterions are valid. The average error rate of cross-validation is 5.477%. Because the goodness of fitting we selected is 0.8, it can be observed that the error rate is relatively high. However, if the error rate is closer to 0%, it might be overfitting and will not be useful for inference.

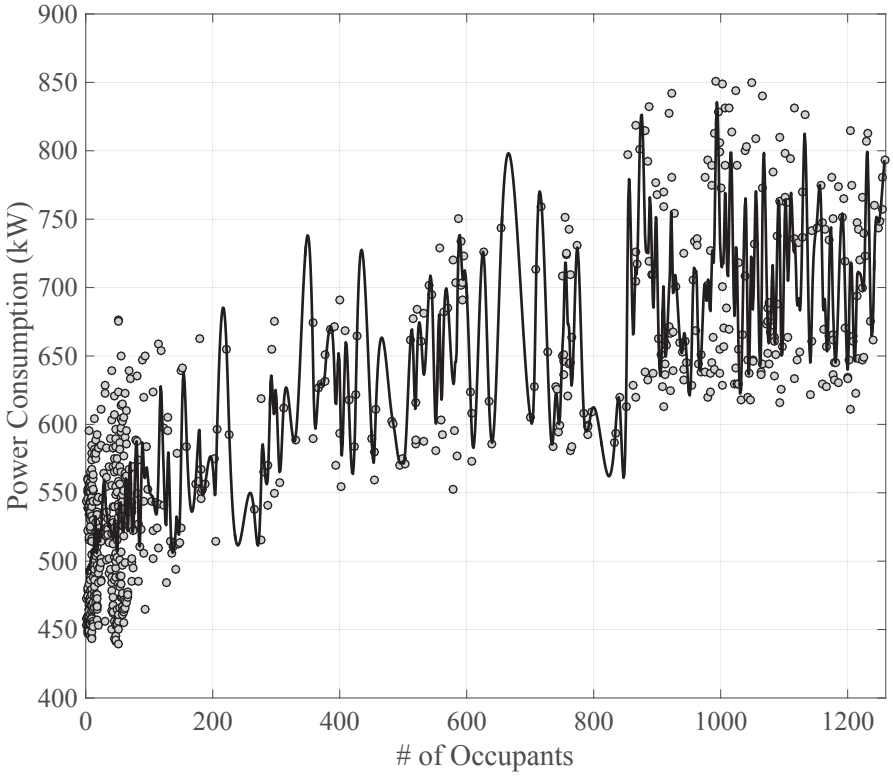

Figure 5.5: Heuristic regression analysis of sample data in the lumped loads.

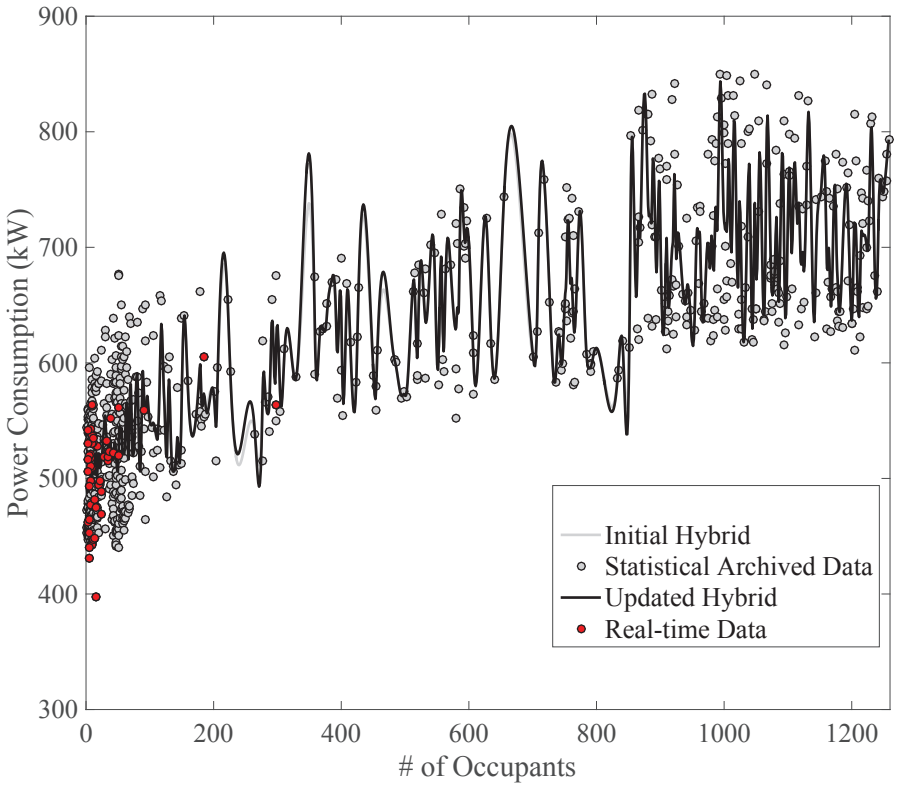

Figure 5.6: Updated hybrid regression model in the test lumped loads.

5.4.2.3 Ongoing Training and Study Results

In this case, we prefer to estimate the power consumption at 8:00 AM October 9 (next Monday). In order to update the parameters in the initial hybrid regression model, a set of real-time data is generated from 12:00 AM to 7:50 AM on October 9. Fig. 5.6 shows the updated hybrid regression model in 2D representation. As illustrated in Fig. 5.6, the structure of the hybrid model remains unchanged while the parameters are slightly updated following the real-time data input.

Table 5.2 shows the error rate analysis between the 2D heuristic model, the 3D

Table 5.2
Error rate analysis of the test lumped loads in a concentrated interval.

Name	OCC. (#)	Time Index	Power Consumption (kW)	Error Rate (%)
Metered	719	49	566.9	-
2D-Hybrid	719	49	581.809	2.63
3D-Model	719	49	578.011	1.96

regression model, and the metering value from the following time index. The model selects the time index as 49 (8:00 AM) with around 719 occupants. The error rate of the 2D hybrid model is 2.63% while the 3D model is 1.96%. Reference to the cross-validation result and the power consumption inference range, I_r, the error rates are in an acceptable range. The error includes the statistical error in data collection and the error might be caused by other factors such as abnormal weather or special events. Predictably, more sample data (monthly or seasonally, not yearly because the influence factors such as climate are distinct) and more uncertain factors considered can construct a more accurate regression model.

Chapter 6

Switching Reconfiguration for Anomaly Detection

6.1 Introduction

Data from utility meters (gas, electricity, water) is a rich source of information for distribution companies beyond billing. The metering infrastructure has further extended from feeder head of a substation throughout the entire feeder loads. Each load should be installed with an electronic meter to constantly observe the energy consumption and to report consumer behaviors through SCADA in the near future. The IP-based communication infrastructure has been gradually improved over time

by extending single flow of information using automated metering reading (AMR) to bi-directional information flow using AMI [95, 120]. Although AMI has improved system reliability and observability within a feeder, application of digital smart meters and addition of a cyber layer to the metering system introduce numerous new vectors for energy stealing caused by meter tampering. While traditional mechanical meters can only be compromised through physical tampering, in AMI, both local and remote data corruption may occur; the metering data can be tampered before being sent to the smart meters, inside the smart meters, and over the communication links [129, 130].

Electrical energy stealing is a notorious problem in electric power systems and has serious implications for both utility companies and legitimate users. The U.S., it is estimated that utility companies lose billions of dollars in revenue every year [47, 131, 132], while in developing countries energy theft can amount to 50% of the total energy delivered [46]. Energy stealing also leads to excessive energy consumption which may cause equipment malfunction or damage [47], and often enables criminal activities, such as the illegal production of substances.

Anomaly detection can be based on two references: sensors and profile. Sensors are employed to monitor and supervise meters in customers' level, the distribution grid, and the communication network. However, the sensor system presents high deployment and maintenance costs, and false alarms may happen during the detection

process [133]. The anomaly detection based on profile analyzes significant variations in consumption patterns [134]. In [135], the feeder remote terminal unit (FRTU) helps narrow the search zone of energy stealing in smart home and community. In [136], the authors proposed a state estimation algorithm utilizing the Kalman filter to find measurement biases for energy usage abnormal detection. Also, [129] compares the real values with the predicted values and apply statistical measures to detect potential fraudulent activities and [137] detects anomalous meter readings on the basis of models built using machine learning techniques on past data. A support vector machine (SVM) classifier was trained in [130, 138] using consumption reports from areas with a high probability of tampering. In addition, the anomaly detection in electrical energy consumers using data mining was proposed in [139] and rough sets were presented in [140].

The strategy of distribution system switch reconfiguration is utilized for fault localization, isolation, system restoration, and optimizing the performance during those procedures. In [141, 142], the authors deal with the problem of optimal reconfiguration of radial distribution networks for minimum loss operations while [143, 144] proposed an optimal strategy for switching devices considers costs of customer outage, installation, as well as annual operation and maintenance. A switch reconfiguration procedure was presented in [145, 146, 147, 148] for distribution system restoration with maximizing the restored load and minimizing the number of switching operations and the analyses in [145, 146] were based on graph theory. The main contribution of this

paper is to combine the profile-based anomaly detection method with switching operations of topology reconfiguration to localize the potential tampered load [149]. A turnable threshold of anomaly will be applied to find the best balance between true and false alarm rates. The line section being detected does not need to be isolated. A graph-theoretic strategy for tampering detection based on profile comparison using reconfiguration switching schemes is presented.

6.2 Profile-Based Anomaly Detection

The profile-based anomaly detection seeks to detect irregularity of potentially tampered data by analyzing abnormal variations in short period consumption patterns. A common assumption in all related works is the deployment of AMI or other real-time meter reading method [83] is performed in the test region. The input dataset to construct the consumption pattern can be obtained from the archived data of from the previous short period that is sufficient to detect ongoing tampering.

An assumption is that offsetting tampers are not performed concurrently, e.g., customer A's consumption is raised while customer B's consumption is lowered. Under this condition, if the actual electricity usage displayed on the central billing center is significantly higher or lower than the meter reading, anomaly (tampering) is suggested. However, the number or amount of tampering in the whole grid is limited, it

is inefficient to compare consumption patterns of all meters on the customer's side. A consumption comparison between the value displayed on the branch head and the summation of all smart meter readings in its downstream can be utilized to localize the tampered lumped loads with at least one tampering. Fig. 6.1 shows two situations with no tampering and at least one tampering in the branch. For the normal one, the summation of meters' reading (red curve) matches the actual usage (blue curve). For the other one, the actual usage curve is higher than the summation of meters' reading (red curve).

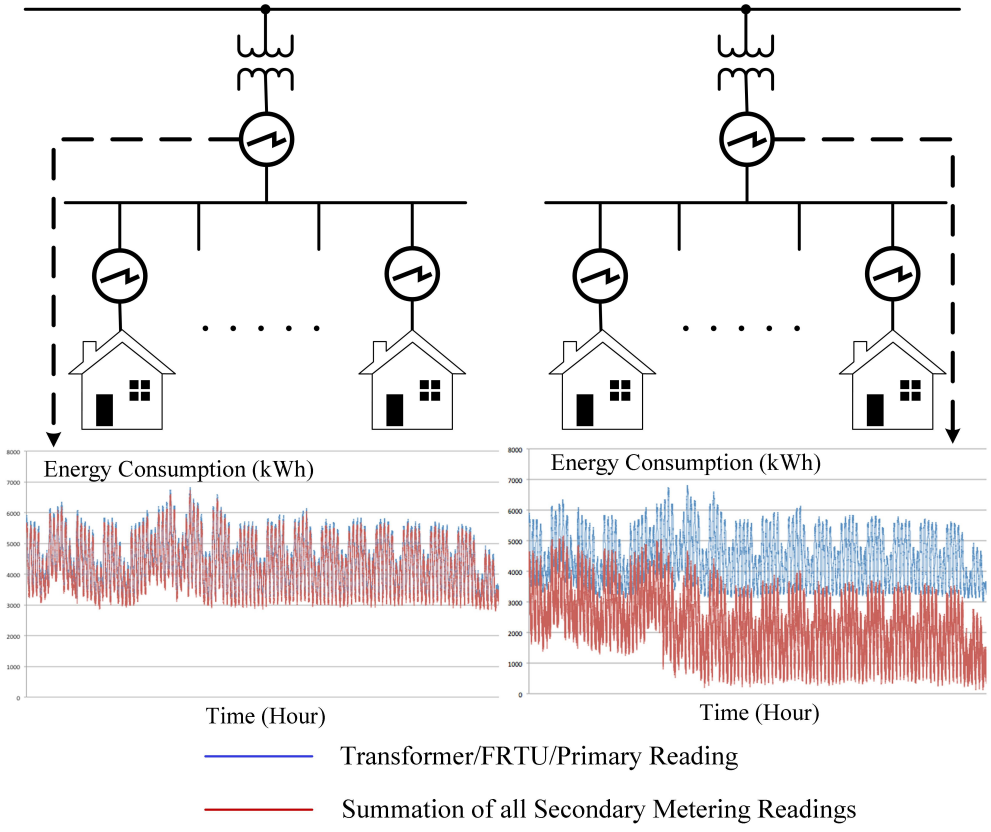

Figure 6.1: An example of a normal smart meter vs. a tampered one.

6.2.1 Threshold of Anomalies

The anomaly threshold of a smart meter identifies whether if a specific reading is anomalous. It compares the current consumption curve with the pattern generated from the previous short period. If the difference between two patterns is greater than the user-defined threshold, the reading is anomalous.

For the comparison between the energy supplied by the grid and the energy reported by smart meters, (6.1) shows the difference of these two in a subsystem during time interval t,

$$\Delta\varepsilon(t) = \varepsilon_{\text{head}}(t) - \varepsilon_{\text{loss}}(t) - \sum \varepsilon_{\text{user}}(t) \tag{6.1}$$

where $\varepsilon_{\text{head}}(t)$ is the energy reported by the low-voltage grid meter (transformer, FRTU, or the primary reading), $\varepsilon_{\text{loss}}(t)$ is the technical losses, and $\varepsilon_{\text{user}}(t)$ is the consumption report of the smart meter from the user side. An abnormal state is triggered when $\Delta\varepsilon(t)$ exceeds the user-defined threshold.

The cost of the tampering detection procedure balanced versus the fraudulent cost

can consider another threshold to decide whether or not to execute the following switching procedure to localize the anomaly. In (6.2), $p_r(t)$ is the real-time price in the subsystem and $\Delta c(t)$ is the fraudulent cost.

$$\Delta c(t) = \sum p_r(t) \cdot \Delta \varepsilon(t). \tag{6.2}$$

6.2.2 Tampered Frequency

The tampered frequency of a smart meter is the ratio between the number of anomalous readings and the total number of readings over a test period. Assume a smart meter obtains the consumption information every τ in minutes, the tampered frequency f_F is shown in (6.3):

$$f_F = \frac{N_{\mathrm{a}}}{60/\tau \cdot 24 \cdot N_{\mathrm{d}}} \tag{6.3}$$

where N_a is the number of anomalous readings and N_d is the number of days during

the test period. For example, assume that a typical smart meter sends out readings every 10 minutes. Utility companies collect one month (30 days) data and 100 readings during this interval are anomalous, then the tampered frequency is around 2.3%. This frequency will be considered at the beginning of the switching procedure to improve the searching efficiency.

6.3 Switching Strategies for Anomaly Detection

The distribution grid can be modeled by a graph $G = (V, E)$ a set of vertices V and a set of edges E. In $G = (V, E)$, substations, distributed energy resources, and distribution loads are referred to as vertices V. The overhead or underground lines with switches are treated as edges E.

6.3.1 Conversion of Distribution Network to a Spanning Tree

A distribution network is constructed by interconnected distribution feeders and microgrids which can be modeled as virtual feeders in this study [150]. The feeders are connected with each other through normally open tie switches. In the process of representing the distribution network using a spanning tree, the vertices denote substations or distributed energy resources are referred to as root nodes and lumped

all root nodes together into one source node in the spanning tree. Since the tie switch is installed in the feeder line section that coupling between the source node and distribution loads, all vertices are connected through the openable edges in the spanning tree that does not contain any loop. An example of distribution network with four sources and the corresponding spanning trees is shown in Fig. 6.2 where each subfigure shows a possibility of reconfiguration. The conversion of the distribution network to a spanning tree is the preliminary to perform the depth-first search (DFS) or other relevant graph-theoretic strategies for switching schemes.

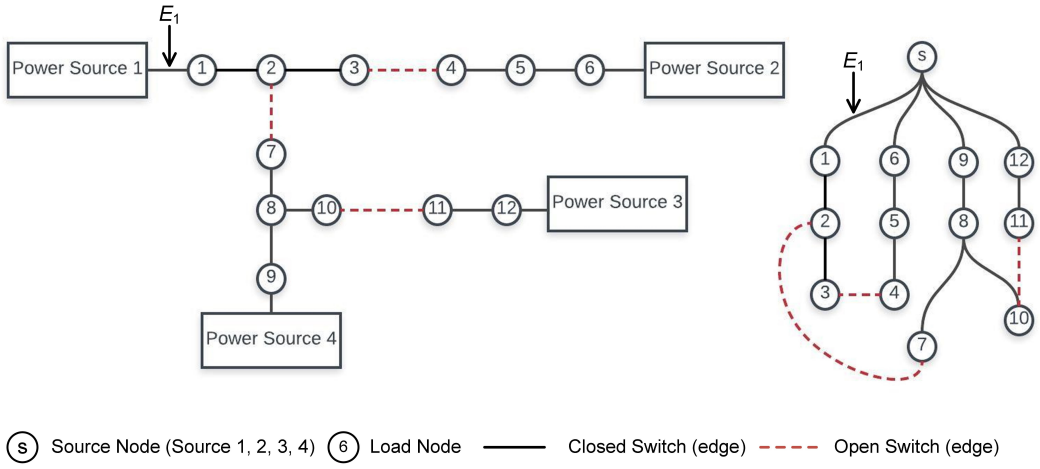

Figure 6.2: Spanning trees of an example distribution network.

6.3.2 Convert Topology Graph to Adjacency or Incidence Matrix

The adjacency matrix is a square $|V| \times |V|$ symmetric matrix that reflects the adjacencies between vertices in a graph while the incidence matrix is a $|V| \times |E|$ matrix that displays vertices and edges incidences [151]. These two matrices can be converted to each other. Since the adjacency or incidence matrix can display the straightforward interrelationship between vertices and edges, the matrix is employed to store and fix the current connection state of the graph topology. Furthermore, this process converts the graph model to the mathematic model which is machine recognizable. The further relevant algorithm and analysis can be performed based on the matrix.

The Algorithm 2 is the pseudocode that shows the procedure that how the graph converts to adjacency matrix. This algorithm can be used not only for a simple virtual network but also a large and complex geographical related topology. The algorithm is summarized as follows:

1. Require $G = (V, E)$ where V is set of vertices and E is the set of edges;

2. Create a pairing function f_{pair} which can uniquely encode two natural numbers into a single natural number [152]. This function will take the x and y coordinate individually and map it into a unique number that will not overlap with other

xy-coordinate of combinations;

3. Make each edge's start vertex which is close to the power source as the head node and the end vertex as the tail node;

4. Perform f_{pair} for each edge head and tail nodes;

5. Combine head and tail nodes together and unique the data string;

6. Renew the head and tail datasets with corresponding nodes indexes;

7. Initialize an empty adjacency matrix;

8. Check each edge in the graph, if the head index is not equal to the tail index, indicate 1 as a connected state to the corresponding position in the adjacency matrix;

9. After generating the adjacency matrix, convert it to the incidence matrix with the existing method [153, 154];

10. Return the adjacency and incidence matrices.

Fig. 6.3 illustrates the sparsity visualization of the adjacency and incidence matrices for the example topology. The adjacency matrix is symmetric (13×13) while the incidence matrix is not (13×12). The E_1 shown in Fig. 6.2 can be represented in the matrix as a dot in Fig. 6.3. Since the graph is undirected, the nonzero elements show the double number of connections in the spanning tree.

Algorithm 2 Graph to Matrix Conversion Algorithm

Input:
$G = (V, E)$
Output:
Creat a pairing function f_{pair}.
`Head` = $f_{\text{pair}}(x_{\text{head}}, y_{\text{head}})$
`Tail` = $f_{\text{pair}}(x_{\text{tail}}, y_{\text{tail}})$
`Node` = unique ([`Head`; `Tail`]).
`Head` = find (`Head` = `Node`).
`Tail` = find (`Tail` = `Node`).
% Initialization. Make sure this is a square matrix.
`AdjM` = zeros (length(`Head`), length(`Head`)).
for `i` = 1:length(`Head`) **do**
  if `Head(i)` $\neq$ `Tail(i)` **then**
    `AdjM` (`Head(i), Tail(i)`) = 1.
    `AdjM (Tail(i), Head(i))` = 1.
  else
    `AdjM (Head(i), Tail(i))` = 0.
  end if
end for
Convert `AdjM` to `IncM`.
return `AdjM, IncM`.

6.3.3 Anomaly Inference with Switching Strategies

The Algorithm 3 is the pseudocode that shows the iterative procedure how the node with tampered load(s) is localized for each iteration. The following variables are defined as:

M_i Incidence matrix representing system topology. If vertex i is connected to edge j, then $M_i[i, j] = 1$, otherwise, 0.

M_a Adjacent matrix representing system topology. If vertex i is adjacent to vertex

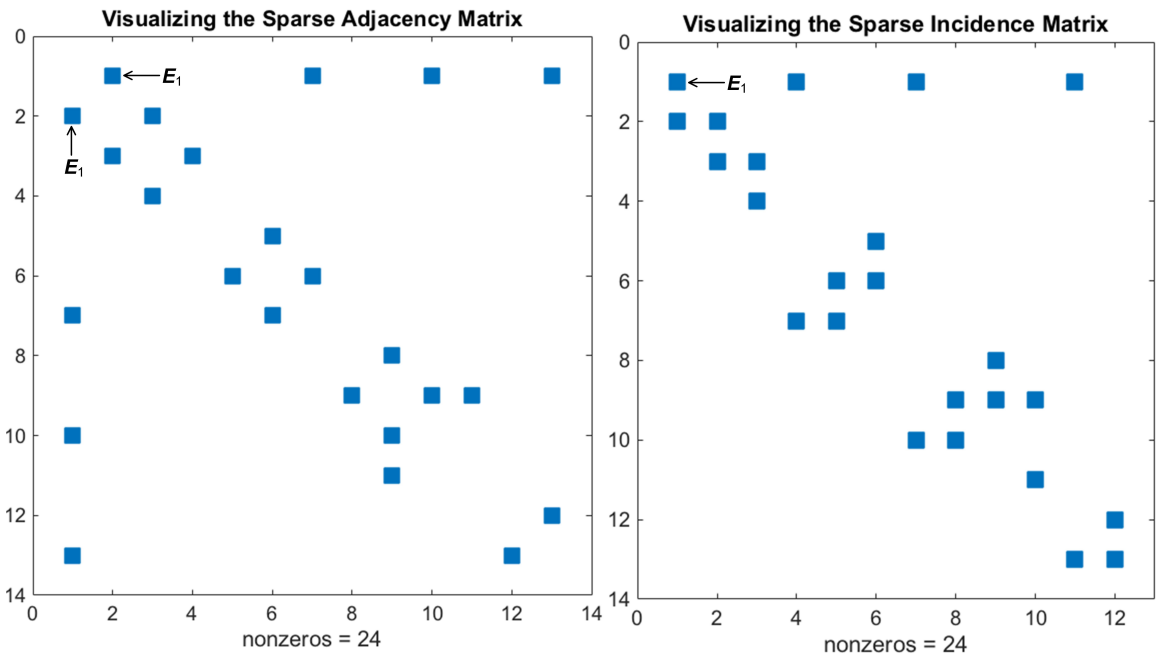

Figure 6.3: Sparsity visualization of the adjacency and incidence matrices for the example topology.

j, then $M_a[i, j] = 1$, otherwise, 0.

V_r Row vector indicating the open/close of switches/breakers/reclosers with length $|E|$. If switch i is open $V_r[i] = 0$, otherwise, 1.

V_s Row vector indicating the location of power sources with length $|V|$. If vertex i connected with a power source, then $V_s[i] = 1$, otherwise, 0.

I_f Index of RTU/FRTU that indicates there exists tampered load(s).

V_g Row vector indicating the location of distributed generators with length $|V|$. If vertex i connected with a distributed generators, then $V_g[i] = 1$, otherwise, 0.

Algorithm 3 Tampered Node Localization Algorithm

Input:

M_i, V_r, V_s

Iteration Fuction: $F_f\ (M_i, V_r, V_s)$

1: $M_i \leftarrow M_i \cdot V_r^T$.
2: $M_a \leftarrow M_i \cdot M_i^T$.
3: Replace all diagonal elements of M_a with 0.
4: % Initialization.
5: $V_f \leftarrow V_s$.
6: $\widehat{V}_f \leftarrow V_f \cdot 0$.
7: % Criteria that if the state is same as the previous iteration.
8: **while** $V_f - \widehat{V}_f \neq 0$ **do**
9: $\quad \widehat{V}_f \leftarrow V_f$.
10: $\quad V_f \leftarrow \widehat{V}_f \cdot M_a + \widehat{V}_f$.
11: **end while**
12: Replace all nonzero elements in V_f with 1.
13: **return** V_f.
14: % Localize the fraud.
15: Find all 0 elements in V_f.
16: **while** More than one tampered nodes in one feeder, change connection states of switches to generate a new V_r. **do**
17: $\quad$ Repeat $F_f\ (M_a, V_r, V_s)$.
18: **end while**

The algorithm proposed here process according to the graph-theoretic switching strategies. A preassumption is a smart meter or tampering indicator is installed in each feeder head. The algorithm starts with balancing the benefits of detecting the tampering and then checking the node with the highest fraudulent frequency in the fraudulent feeder first. The DFS is then performed to decide the following node to be checked. Since the incidence matrix is constructed by vertices and edges, the changing of connection status can be displayed on this matrix directly, the input topology in the Algorithm 3 is the incidence matrix. The algorithm is summarized as follows:

1. Convert the incidence matrix to the adjacency matrix with current connection status of switches for fraudulent nodes detection;

2. Initialize V_f with V_s and set $\widehat{V}_f$ as the previous iteration result of V_f and initialize $\widehat{V}_f$ with 0;

3. Check every nodes status of each iteration by checking $V_f = \widehat{V}_f$;

4. If status of all nodes are stable, replace all nonzero elements in V_f with 1;

5. For the situation that more than one fraudulent nodes in one feeder, change connection states of switches to generate a new V_r and repeat F_f (M_i, V_r, V_s);

6. Return V_f;

7. Find 0 element(s) in V_f to localize the tampering node.

6.4 Anomaly Detection Analysis for Case Study

This section is to validate the proposed graph-theoretic strategy for anomaly detection using reconfiguration switching schemes with a topology of a distribution system. The procedures of changing connection status of switches to reconfigure the topology and the detection results of potential attacked load nodes in each iteration will be discussed in this section.

6.4.1 Test Case Description

Fig. 6.4 is the topology of the distribution test network. There are two substations, four source nodes, and 20 load nodes. The four switches at each feeder head are deployed with a anomaly indicator or a smart meter to perform the profile-based anomaly detection. The four load nodes with the highest tampered frequency at each feeder are indicated by red boxes and the randomly generated tampered load node is in a green circle. The corresponding spanning tree is shown in Fig. 6.5. The four source nodes are lumped into a single source and the indicator installed in E_4 infers an anomaly that may be subject to anomaly.

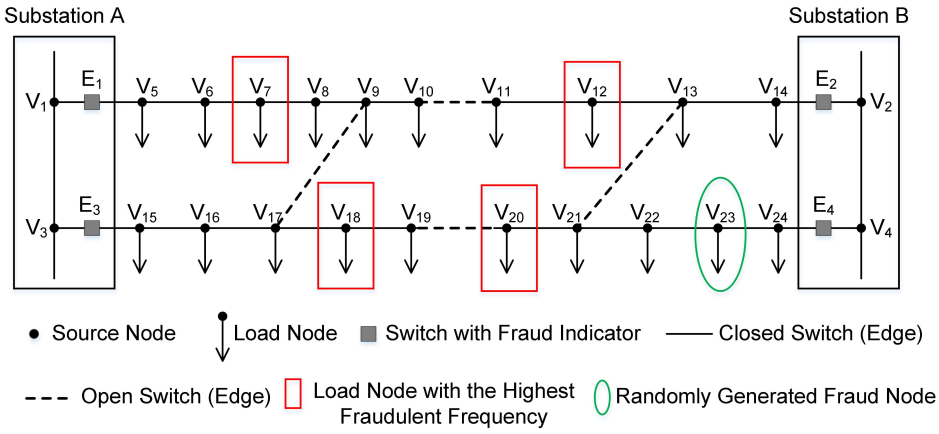

Figure 6.4: Topology of distribution test network.

Fig. 6.6 illustrates the sparsity visualization of the adjacency and incidence matrices for the case topology. The adjacency matrix is symmetric (21×21) while the incidence matrix is not (21×20). As mentioned in Section III, Part B, the source node

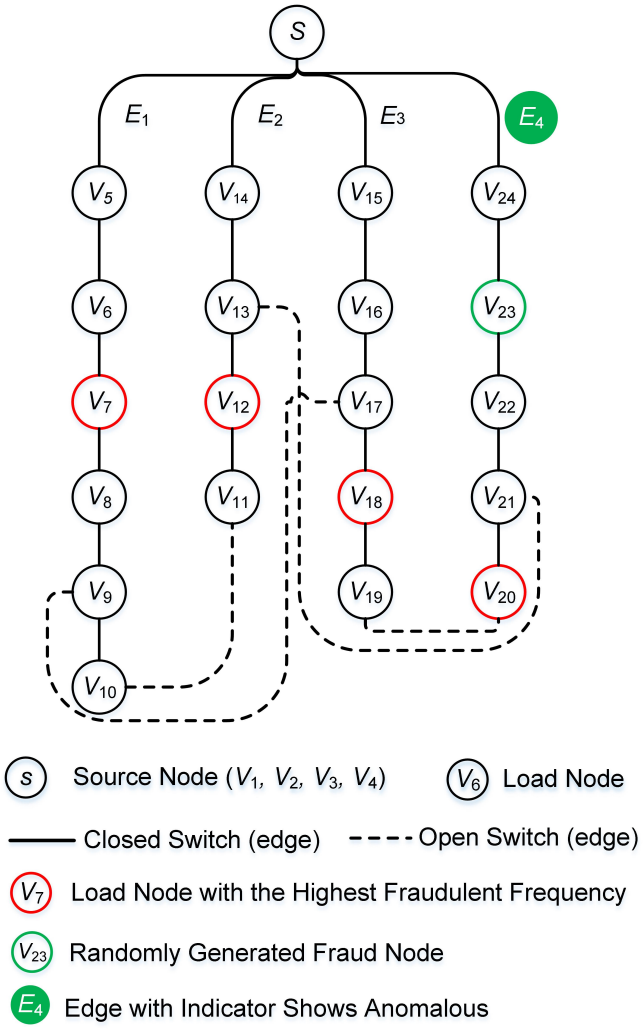

Figure 6.5: Spanning tree of a case.

S is treated as the node with index 1 while vertex 5 (V_5) as node 2, vertex 6 (V_6) as node 3, and so on. The corresponding position of E_4 has been indicated in Fig. 6.6. Since the graph is undirected, the nonzero elements show the double number of connections in the spanning tree.

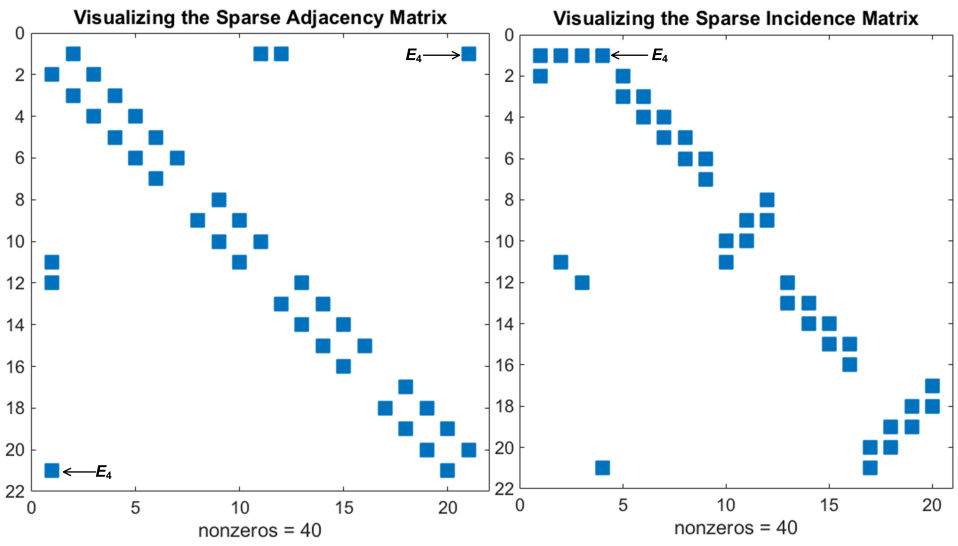

Figure 6.6: Sparsity visualization of the adjacency and incidence matrices for the case topology.

6.4.2 Switching Procedures

In all steps, since all source nodes are lumped as node S and it is the only source node, the power source vector $V_s = [1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$.

Step 1: Since V_{20} has the highest tampered frequency, the switching procedure starts from checking the state of V_{20}. As shown in Fig. 6.7 (a), we broke the switch between V_{20} and V_{21} and closed the switch between V_{19} and V_{20} simultaneously. Since the current connection status of the topology has been changed, we modified values on corresponding positions in the input incidence matrix. The consumption reading of the indicator installed in E_3 displayed a corresponding increase but within a normal range. The E_4 still showed anomalous means the tampering was in the load node V_{21},

V_{22}, V_{23}, or V_{24}. The value of switch E_4 in V_r is 0.

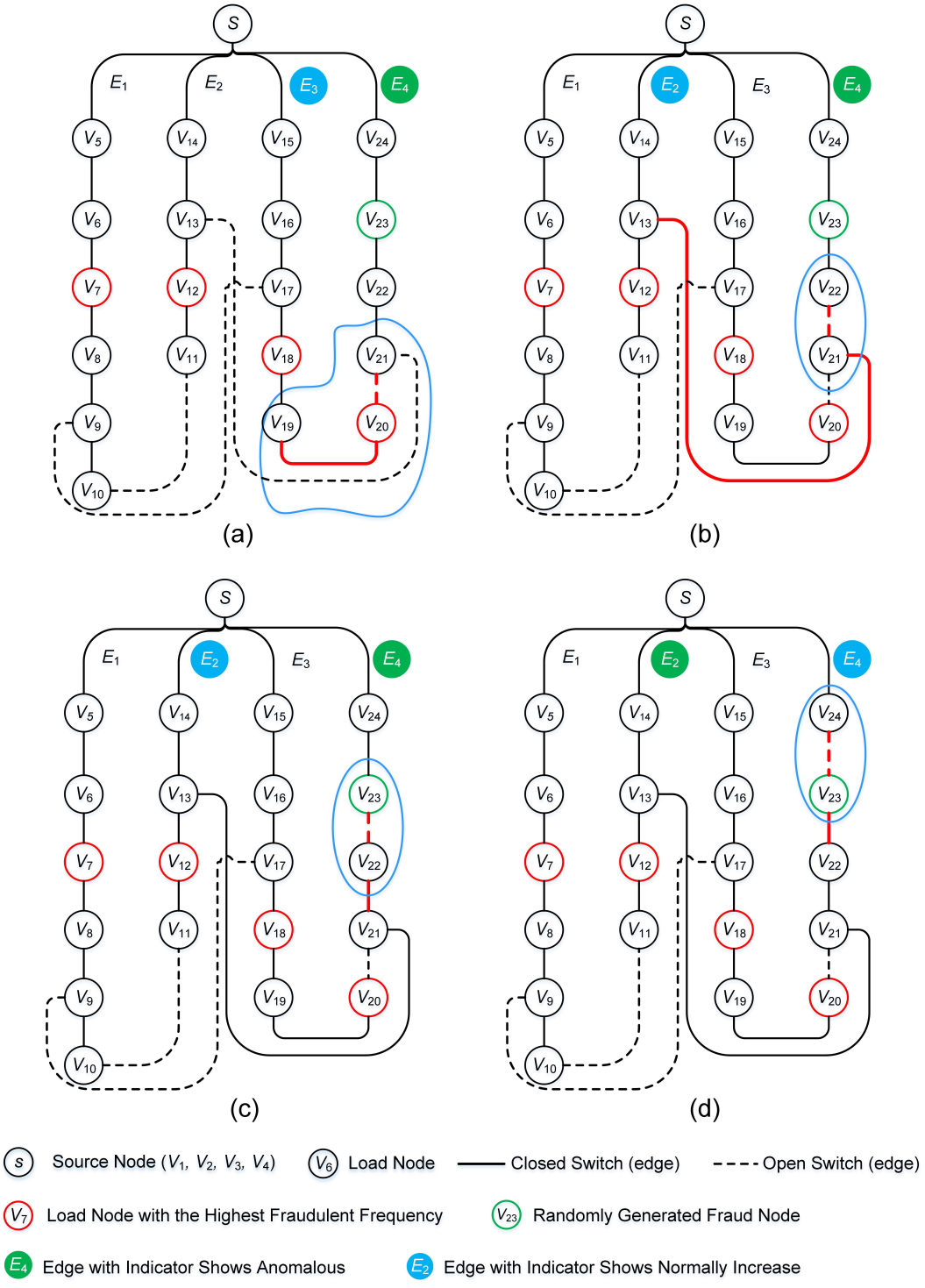

Figure 6.7: Graph representation of switching procedures.

Step 2: This step is displayed in Fig. 6.7 (b). According to DFS, we then broke the switch between V_{21} and V_{22} and closed the switch between V_{13} and V_{21} as shown in the blue circle to check the state of V_{21}. Similiar like in Step 1, the consumption reading in E_2 raised due to the additional load in V_{21} while the anomalous reading in E_4 indicated the tampering existed in V_{22}, V_{23}, or V_{24}.

Step 3: As demonstrated in Fig. 6.7 (c), the search is following the connection sequence within the abnormal feeder and this step is to check the state of V_{22}. Repeat the switching operation to open the switch between V_{22} and V_{23} and closed the switch between V_{21} and V_{22}. The data report obtained from E_2 feeder added up the load consumption data from V_{22} but within an acceptable limit. The tampering was detected in either V_{23} or V_{24}.

Step 4: Repeat the switching procedure as previous steps. As illustrated in Fig. 6.7 (d), we opened the switch between V_{23} and V_{24} and closed the switch between V_{22} and V_{23} in this step. As a result, the anomaly indicator installed in E_2 responded an alarm since the tampered load node V_{23} was connected to this feeder and the consumption reading in E_4 seemed normal. Therefore, we can decide that the tampering was in the load node V_{23}.

6.4.3 Customer-Level Tampering Detection

After localizing the tampered load node, the profile-based anomaly detection using consumption patterns comparison between the previous and the current period was performed to find the tampered meter on customers' level. There might be more than one abnormal meters on this node.

Chapter 7

Anomaly Node Searching Incorporating Distributed Generation

7.1 Introduction

There are two types of losses in power systems: technical and non-technical [155]. Technical losses are consisting mainly of inevitable power dissipation in power generation, transmission, and distribution while the non-technical losses consist of electricity theft, non-payment energy usage, and errors in recording and billing. As the most

egregious factor, the amount of energy theft is on the rise. Up to $6 billion of electricity is stolen in the U.S. annually [47, 114], while in developing countries, energy theft can amount to 50% of the total energy delivered [46]. The traditional mechanical meters can be compromised through physical tampering. Although the advanced metering infrastructure (AMI) has improved system reliability and observability to measure and control energy usage in communicating through metering devices, the development of the AMI network brings with it security issues, including tampering metering data [129, 130] and the increasingly serious risk of malware in the new emerging network.

The metering data can be tampered before being sent to the smart meters, inside the smart meters, and over the communication links while the malware is usually embedded in the data payloads of legitimate metering data. In time-critical communications, it is difficult to detect malware in metering devices, which are resource-constrainted embedded systems. In 2009, Mike Davis at Blackhat demonstrated the speed at which smart meters could be compromised by malware and how quickly malware could propagate worms in the meters [156]. In 2010, the Stuxnet worm [42] attacked SCADA (Supervisory Control And Data Acquisition) systems and PLCs (Programmable Logic Controllers) in industrial systems [157, 158]. Different than conventional malicious attacks, several smart meters might share a same identity (ID) or one physical device may generate an arbitrary number of an additional ID, which can launch Distributed Denial-of-service attack (DDoS) attacks. DDoS only

affects the unavailability of smart meters. However, this attack can be utilized to reduce the share of resources in the topology and give attackers more information to perform other attacks [159, 160]. A McAfee report warned that an attacker could exploit smart meters easily and takes control of the whole system [161]. Since the system specifications, diverse network protocols, and operational-constrainted metering devices in AMI, the existing defense methods against malicious attacks cannot be applied directly. The main motivation of this work is not to provide a defend method to improve the cyber security of the grid network. It proposed a switching scheme to detect and localize the occurred attacks or tampering behaviors within a distribution system.

7.2 Related Works

Researchers to prevent abnormal electricity usages have done many existing methods. Machine learning and data mining [139] have been widely applied for extensive intelligent data analysis to recognize normal consumption patterns such that deviations can be detected as anomalies. Authors in [137] using machine learning techniques on the past data to built models for detecting anomalous meter readings. A technique can identify diverse forms of tampered activities by using artificial neural networks and smart meter fine-grained data in [162]. A state estimation algorithm utilizing the Kalman filter to find measurement biases is proposed in [136] while a support vector

machine (SVM) classifier was trained in [130, 138] using consumption reports from areas with a high probability of energy theft. Another main strategy of detecting malicious meters is based on the real-time comparison. In [135], the FRTU helps narrow the search zone of energy fraud in smart home and community. Also, [129] compares the real values with the predicted values and apply statistical measures to detect potential tampered activities. The methods to distinguish malware-bearing traffic and legitimate metering data using a disassembler and statistical analysis are described in [163, 164]. Different than data based detection approaches, [165, 166] provide the attack and intrusion detection mechanism for a smart grid neighborhood area network (NAN). In addition, some strategies based on graph theory have been developed. In [167], a novel inspection algorithm identifies each meter with a unique binary-coded number to locate the unique malicious meter is proposed. The conversion between the power grid and binary tree or spanning tree is adopted in [130, 168, 169]. Sensors can also be employed to monitor and supervise meters in customers level, the distribution grid, and the communication network. However, the sensor system presents high deployment and maintenance costs, and false alarms may happen during the detection process [133].

The reconfiguration switching scheme in the distribution system is mainly utilized for fault localization, isolation, system restoration, and optimizing the performance during those procedures. The main purpose to exploit this problem is to achieve

optimization. In [141, 142], the authors deal with the problem existed in radial distribution networks for minimum loss operations while [143, 144] proposed a strategy for switching devices considers costs of customer outage, installation, as well as annual operation and maintenance. In [145, 146, 147, 148] introduce methods with maximizing the restored load and minimizing the number of switching operations to achieve distribution system restoration based on the graph theory.

7.3 Switching Strategies for Tampered Node Localization

In this section, the reconfiguration switching scheme and the profile-based tampering detection are combined to localize the subfeeder, microgrid, cluster, or the distribution node with tampered load(s) or malicious meter(s). Based on the economic consideration, the non-technical losses should be larger than 5% of the normal power consumption. All electrical utilities keep the power on during the switching procedures while meeting electrical and operational constraints.

7.3.1 Convert the Distribution Network To a Graph

A distribution grid is constructed by feeders and microgrids [170]. The feeders are connected with each other through normally open tie switches. According to the conversion process mentioned in Chapter 6, an example of a distribution network with two substations and the corresponding adjacency and incidence matrices are shown in Fig. 7.1. The sparsity visualization of the adjacency and incidence matrices for the example topology are also illustrated in Fig. 7.1. It should be noted that the order of vertices and edges may change the form of matrices but can not change the topology of the graph. In this example, the order of vertices following the numbering sequence in this figure while the S_1 to S_5 represent edges 2 to 7. The RTU_1 and RTU_2 represent edge 1 and 7, respectively. Two vertices V_3 and V_4 and their connection S_1 (edge 2) are displayed on the visualized sparse matrices. The nonzero elements of these two matrices represent the number of edges in the graph. Since the graph is undirected, each edge is counted twice. This example is used for the case study in this paper.

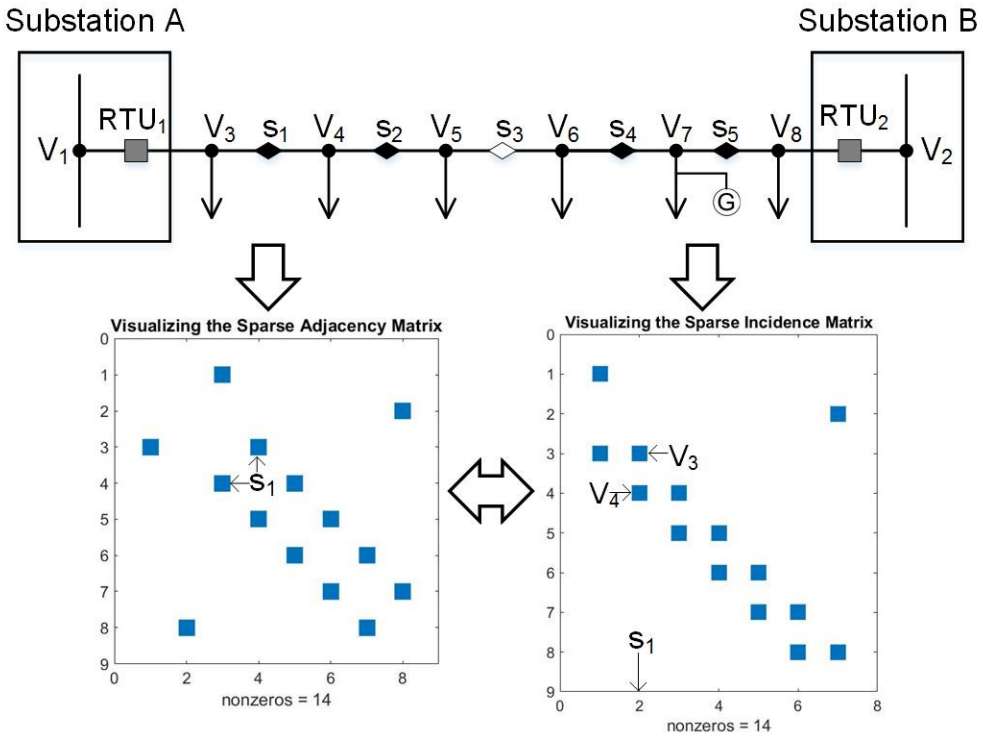

Figure 7.1: Sparsity visualization of the adjacency and incidence matrices for the example topology.

7.3.2 Tampered Node Localization with Switching Strategies

The Algorithm 4 is the pseudocode that shows the more effective iterative procedure how the node with tampered load(s) is localized for each iteration cooperated with distributed generators. The following variables are defined as:

M_i Incidence matrix representing system topology. If vertex i is connected to edge j, then $M_i[i, j] = 1$, otherwise, 0.

M_a Adjacent matrix representing system topology. If vertex i is adjacent to vertex

j, then $M_a[i, j] = 1$, otherwise, 0.

V_r Row vector indicating the open/close of switches/breakers/reclosers with length $|E|$. If switch i is open $V_r[i] = 0$, otherwise, 1.

V_s Row vector indicating the location of power sources with length $|V|$. If vertex i connected with a power source, then $V_s[i] = 1$, otherwise, 0.

I_f Index of RTU/FRTU that indicates there exists tampered load(s).

V_g Row vector indicating the location of distributed generators with length $|V|$. If vertex i connected with a distributed generators, then $V_g[i] = 1$, otherwise, 0.

The algorithm starts with balancing the benefits of detecting the tampered node, only the detected power losses are more than 5%, the process is performed. The algorithm is summarized as follows:

1. Isolate all subfeeders or clusters with distributed generators, if I_f shows no tampered load(s), the tampered load(s) exists in the isolated subfeeders or clusters.

2. Restore the connection of each microgrid sequentially and check I_f to identify which microgrid has the tampered load(s).

3. If the tampered load(s) is not in the isolated area, perform the iteration function to localize the tampered node.

Algorithm 4 Tampered Node Localization Algorithm

Input:

M_i, V_r, V_s, I_f

Isolate all subfeeders or clusters with distributed generators.
if I_f shows no tampered load(s) in this system.
then
  $V_f = V_g$.
  Restore the connection of each subfeeder or cluster sequentially and check I_f.
else
Iteration Fuction: $F_f\ (M_i, V_r, V_s, I_f)$
  $V_r(I_f) = 0$;
  ▷ V_r keeps updating;
  ▷ Convert the incidence matrix to the adjacency matrix with current connection status;
  $M_i \leftarrow M_i \cdot V_r^T$.
  $M_a \leftarrow M_i \cdot M_i^T$.
  Replace all diagonal elements of M_a with 0.
  ▷ Initialization;
  $V_f \leftarrow V_s$.
  $\widehat{V}_f \leftarrow V_f \cdot 0$.
  ▷ Criteria that if the state is same as the previous iteration.
  while $V_f - \widehat{V}_f \neq 0$ **do**
    $\widehat{V}_f \leftarrow V_f$.
    $V_f \leftarrow \widehat{V}_f \cdot M_a + \widehat{V}_f$.
  end while
  Replace all nonzero elements in V_f with 1.
  return V_f.
  ▷ Localize the tampered node.
  Find all 0 elements in V_f.
  while More than one tampered nodes in one feeder, change I_f to generate a new V_r. **do**
    Repeat $F_f\ (M_a, V_r, V_s, I_f)$.
  end while
end if

4. Convert the incidence matrix to the adjacency matrix with current connection status of switches for tampered nodes detection;

5. Initialize V_f with V_s and set $\widehat{V}_f$ as the previous iteration result of V_f and initialize

$\widehat{V}_f$ with 0;

6. Check the status of every node during each iteration by checking $V_f = \widehat{V}_f$;

7. If status of all nodes are stable, replace all nonzero elements in V_f with 1;

8. For the situation that more than one tampered nodes in one feeder, change I_f to generate a new V_r and repeat $F_f\ (M_i, V_r, V_s, I_f)$;

9. Return V_f;

10. Find 0 element(s) in V_f to localize the tampered node.

7.3.3 Customer-Level Fraud Detection

In this study, we assume all customers have smart meters. After localizing the tampered load node, the profile-based tampering detection using consumption patterns comparison between the previous and the current period was performed to find the potential fraud on customers' level. There might be more than one fraud on this node.

7.4 A Case Study

This section is to validate the proposed graph-theoretic strategy for tempering detection using reconfiguration switching schemes with a topology of a distribution system. The procedures of changing connection status of switches to reconfigure the topology keeps the whole system power on and meet the electrical and operational constraints.

7.4.1 Test Case Description

Fig. 7.2 is the topology of the distribution test network. There are two substations and six load nodes. The two switches at each feeder head are deployed with a substation RTU to perform the profile-based tampering detection. The two load nodes with the distributed generators can generate the microgrids [171, 172]. The randomly generated tampered loads are detected by RTU_2.

The corresponding input incidence matrix can be gathered from the previous section. The input row vectors of the first scenario in Fig. 7.2 are as $V_s = [1\ 0\ 0\ 0\ 0\ 0\ 0\ 1]$ since V_1 and V_2 are two power sources; $V_r = [1\ 1\ 1\ 0\ 1\ 1\ 1]$ due to the normally open tie switch S_3; $I_f = 7$ for the random fraud alarm and $V_g = 7$ because of the distributed generator. In the second scenario, the V_s, I_f, and V_g remain unchanged

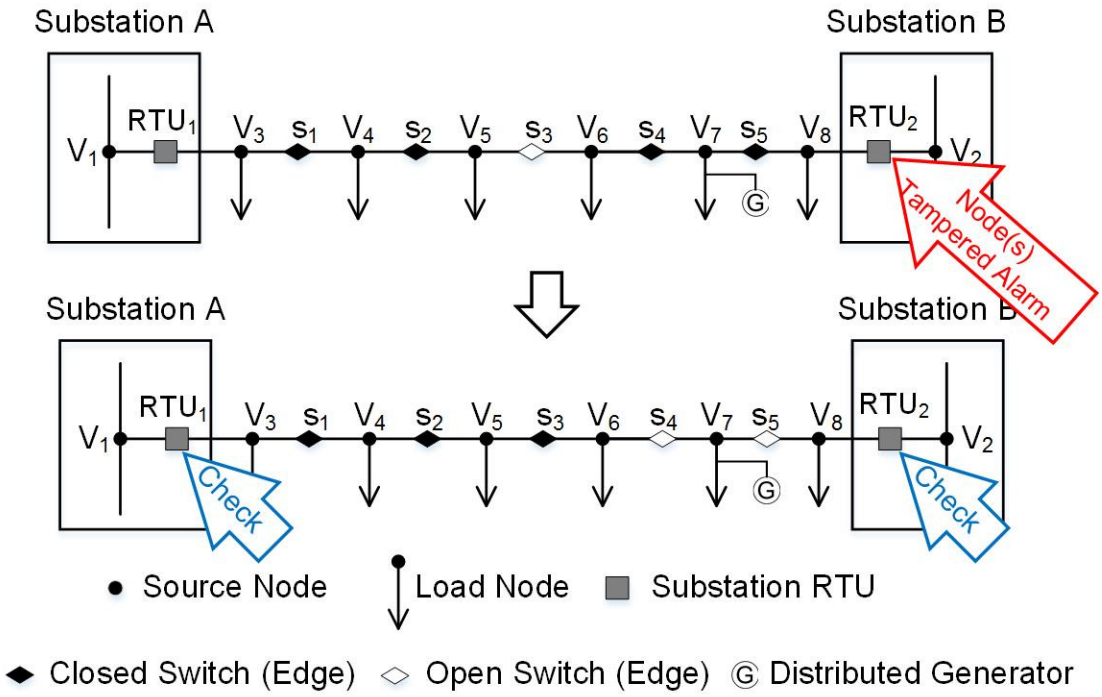

Figure 7.2: Switching procedures to localize the tampered load node in the example topology.

while $V_r = [1\ 1\ 1\ 1\ 0\ 0\ 1]$ since the switching procedure.

7.4.1.1 Switching Procedures

Since the topology of the test system keeps unchanged, the input incidence matrix has no changes. The difference between each iteration is the updated connecting status V_r.

Step 1: Under the operational constraints, connect the tie switch between these two feeders in order to keep the test system power on during this procedure.

Step 2: Isolate the load node V_6 since it connected with a distributed generator which can supply the power for this microgrid.

Step 3: Perform the profile-based tampering detection at RTU_2, if RTU_2 shows normally, the tampered loads are not in V_7.

Step 4: Perform the profile-based tampering detection at RTU_1, if RTU_1 shows normally, the tampered loads are in V_6; If not, the tampered loads are in V_5.

In this work, a simple virtual distribution network is applied, since the main content of this work is in the strategy of switching the openable edges, the search process could be an exponential time search or only spend a linear time in a real distribution network. This is the preliminary work that intend to prove the feasibility of the proposed idea.

Chapter 8

Inference of Massive Tampering

8.1 Introduction

The discrepancy between actual usage and metering information occurs due to illegal or irregular electricity consumers, tampering of meters or records, and unmetered supplies [24, 25]. Among those, large amounts of these losses are absorbed by fraudulent electricity consumers. Up to $6 billion of electricity is stolen in the U.S. annually, while in developing countries, energy theft can amount to 50% of the total energy delivered [46, 114, 155, 173]. Traditional physical tampering includes directly using alternative neutral lines to connect unregistered electrical appliances, sabotaging control wires and feeders, using magnets to decelerate the spinning discs for recording

the energy consumption, and tapping off of a neighbor/legal consumer [24, 155, 174]. In order to improve the system reliability and observability to measure and control energy usage in communicating through IP-based metering devices, utilities around the world have been refining their business models to provide additional choices to consumers by deploying advanced metering infrastructure (AMI). However, the development of the AMI network also brings with it security issues, including tampering metering data [129, 130, 175, 176, 177] and the increasingly serious risk of malware in the new emerging network [178]. It is apparent that knowing how to identify cases of anomalous discrepancy accurately is vital for utilities. Such identification provides a means of devising and implementing suitable preventative and corrective means of reducing the losses involved [174].

Cybertampering is an electronic alteration from their real consumption values in metering devices through the cyber attack. The most common situation is the malicious customers may reduce the scale of their energy usages to fake their electricity bills [177, 179]. The massive tampering in a distribution system can utilize vulnerabilities in multiple connections, which include exploitation on existing communication media and protocols, segregation of network architectures, and false data injection, to attack multiple metering devices (usually more than 20%) simultaneously. Malware usually runs in a system through the concealed installation and embedded in the data payloads of legitimate metering data. When the invasion of a single meter is completed, the worm or virus quickly infects nearby meters and can even invade the main server

and host [180, 181, 182], causing massive tampering. The infected devices with malware can be manipulated to perform unauthorized use of services covertly [183]. In 2009, Mike Davis at Blackhat demonstrated the speed at which SMs could be compromised by malware and how quickly malware could propagate worms in the meters [156]. In 2010, the Stuxnet worm [42] attacked SCADA systems and programmable logic controllers (PLCs) in industrial systems [157, 158]. In 2015, due to the attack of the malware named BlackEnergy, the Ukrainian power grid suffered a sudden power outage, causing more than 700,000 households in western Ukraine under power outages for several hours [184]. Although the distributed denial-of-service (DDoS) attack only affects the unavailability of SMs, this attack can be utilized to reduce the share of resources in the topology and give attackers more information to perform other attacks [159, 160]. A McAfee report warned that an attacker could exploit SMs easily and takes control of the whole system [161].

Malware detectors and anti-virus software should be installed in the alarm system of metering devices to prevent intrusion and/or track the footprint of malware [185]. The alarm system of the modern power grid mainly includes anomaly detection of the information network and the operation monitoring of the physical system combining with the abnormal data injection. The intrusion detection system is used in the information network to monitor the communication traffic in the smart grid. When an attack occurs, the intrusion detection system analyzes traffic based on specific rules and generates an alert with an IP/MAC address [186]. In a physical system, data

is mainly the power parameters such as voltage/current amplitude, voltage/current phase, etc. These measurements can be used to estimate the state of the system and send an alarm to the operator if the state is abnormal [74]. Since the communication topology and the physical topology in the smart grid are interrelated as shown in Fig. 8.1, the IP/MAC address and suspicious nodes/lines in the physical system can be used to correlate their respective security factors or trustworthy parameters and be used for system analysis. Also, these parameters can combine with the statistic of anomaly occurs in metering devices, communication networks, and servers/hosts of the system to establish the malware detection model [187, 188] in the alarm system. Therefore, a complete and reliable alarm system can monitor the network and the physical layers at the same time, which can help the control center to get better countermeasures for different types of intrusions.

The main motivation of this work is not to provide a defend or maintain strategy to improve the cybersecurity of the grid network, it proposed a scheme to detect ongoing malware attacks or tampering behaviors within a distribution subsystem. The main contribution of this paper is to propose a massive tampering inference method associated with a probabilistic trust model. The trustworthy computing is based on the statistics of alarms and event logs from the metering system.

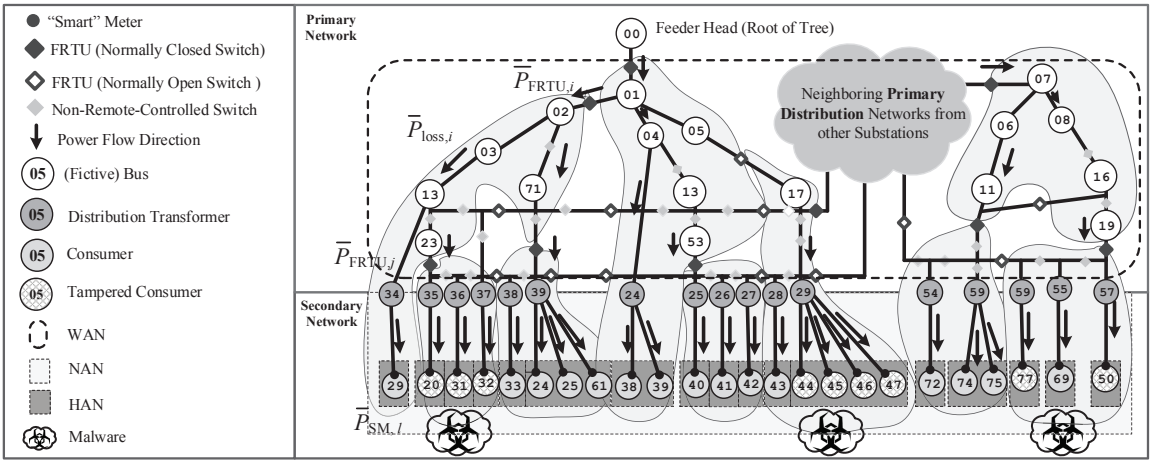

Figure 8.1: An example of an electrical distribution system with massive tampering in subsystems via malware.

8.2 Model Formulation for Massive Tampering Inference

Fig. 8.1 demonstrates an example of an electrical distribution system with massive tampering in subsystems via malware. The power injection starting from the feeder head of a primary distribution feeder to end consumers. With the presence of FRTUs, the distribution feeder can be grouped into several subsystems, e.g., subsystems are partitioned by light gray areas. The home area network (HAN) provides interfaces into the home and business (end-node) for energy consumption monitoring and to support demand response functionality [189]. The HAN ensures the communication from the meter to devices inside the consumers home and allows devices located within a home to communicate with each other. The neighborhood area network

(NAN) provides communication between devices or meters in a small area, typically extending the reach to the majority of the meter population [189]. The data exchange might occur between HANs and NANs but not with the wide area network (WAN). It provides communications from the utility head end out to devices in the field. Field access or collection points on the edge of the WAN provide connections to and/or consolidation of meter data collected by the NAN for retrieval [177, 189]. The WAN is also used to communicate with all or specific devices in the field to initiate tasks, upgrade firmware or request specific data. The worm propagation and malware injection often occur in the HAN/NAN level [165, 166]. Once malware gains installation permissions via the physical and/or electronic layer, it is highly likely to cause a large number of vulnerabilities which will lead to massive tampering. Anomaly detectors and the alarm system may be set up in NAN and the control system [175].

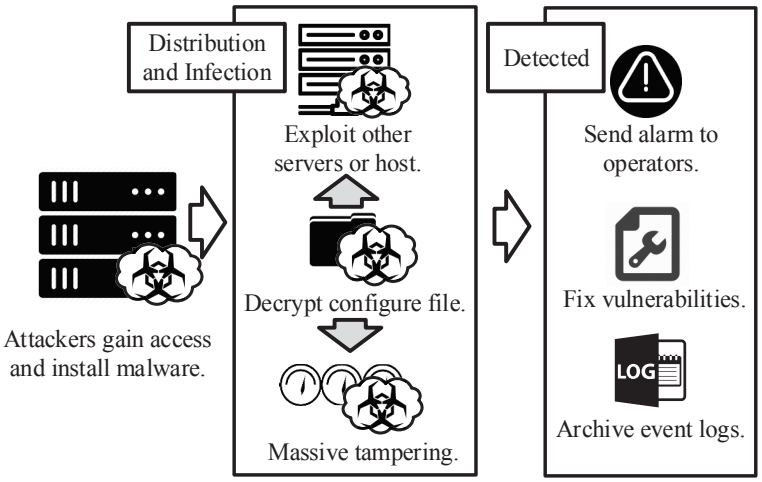

Figure 8.2: Generation of alarm and event log according to the malware detection.

The process of malware attack is illustrated in Fig. 8.2. Attackers gain access through

malicious code (electronic layer) or other artificial means (physical layer) and install malicious software to a server. Malicious code quickly decrypts the configuration file of the target operating platform and creates a large number of vulnerabilities. The malware then spreads viruses or worms to nearby or higher-level servers or meters, causing massive tampering. Once the malware is detected, the corresponding vulnerabilities are fixed and the system sends an alert to the console and generates an event log for future reference. The proposed approach according to the statistics of alarms (not just malware alarm) to generate a probabilistic trust model to evaluate the possible discrepancy.

Ideally, as subsystems illustrated in Fig. 8.1, the measurements from the zonal SCADA system should be equal to the summation of all home SMs' readings with an acceptable power loss. The root FRTU serves as a power injection point of the subsystem, the FRTU(s) located at the points downstream subtrees are treated as the border switches, which are the leaf nodes of a spanning tree. Real-time power flow measurements of these FRTUs represent lumped loads of the other subsystems connected to them. Power losses for the given subsystem can be estimated by applying any distribution power flow method. Due to some reported device failure or data losses (based on alarms), there is an extra energy loss $\Delta E_{t,i}$. The alarm-based trust model is applied for each FRTU and SM to estimate the detected anomalous energy losses. Once $\Delta E_{t,i}$ exceeds the detected variation from alarm reports as described in (8.1), the undetected massive tampering might exist in the subsystem, an anomalous

state is triggered.

$$\Delta E_{t,i} = | \underbrace{(\overline{P}_{\mathrm{FRTU},i} \cdot x \cdot t_{\mathrm{p}} - \sum_{\mathrm{j} \in \mathfrak{J}_i} \alpha_{t,j} \cdot \overline{P}_{\mathrm{FRTU},j} \cdot x \cdot t_{\mathrm{p}})}_{\substack{\text{measurement from} \\ \text{primary network,} \\ \text{i.e.SCADA}}} - \underbrace{\sum_{l \in \mathfrak{L}_i} \beta_{t,l} \cdot \overline{P}_{\mathrm{SM},l} \cdot t_{\mathrm{s}}}_{\substack{\text{measurement from} \\ \text{secondary network,} \\ \text{i.e.smart meters}}} - \overline{P}_{\mathrm{loss},i} \cdot t_{\mathrm{s}}|, \tag{8.1}$$

$$\alpha_{t,j}, \beta_{t,l} \in [0,1],\ t_{\mathrm{p}} = \frac{t_{\mathrm{s}}}{x}.$$

$\Delta E_{t,i}$ detected anomalous energy losses in i-th subsystem during time period t.

$P_{\mathbf{FRTU},i}$

measurement of power consumption by the root FRTU in the i-th subsystem.

$P_{\mathbf{FRTU},j}$

power measurement by the j-th FRTU, which is one of the leaf FRTUs (nodes) in the i-th subsystem.

$P_{\mathbf{SM},l}$ power measurement of l-th home SM in i-th subsystem.

$P^t_{\mathbf{loss},i}$ technical power losses of the i-th subsystem, which can be estimated by applying any distribution power flow analysis.

$t_{\mathbf{s}}$ time cycle of gathering metering measurements from SMs, it is usually set as

15-min.

$\mathfrak{J}_i$ set of FRTUs leaf nodes in the i-th subsystem.

$\mathfrak{L}_i$ set of home SMs belonging to the i-th subsystem.

$\alpha_{t,j}$ trust model based on alarm statistics for the measurement from the primary network in the j-th device in t.

$\beta_{t,l}$ trust model based on alarm statistics for the measurement from the secondary network in the l-th meter in t.

$t_{\mathbf{p}}$ time cycle of gathering measurements from FRTU, it is usually set as 1 to 3 seconds.

x normalization parameter.

A compromised cybersecurity of a SCADA system can cause serious damage to a power system if the attack is able to launch disruptive switching actions leading to a loss of load or equipment damage. This is particularly troublesome if the attack can penetrate the control center network that is connected to distribution substations under the SCADA system [190]. The impact analysis is the task to analyze the intrusion behaviors and the potential threatens, evaluate the resultant consequences of the cyber attacks, and consider the economic balance. The operator should balance the detection cost versus the tampered power loss. Another threshold to decide whether or not to execute the following switching procedure to localize the tampered

area could be considered. In (8.2), Px_t is the real-time price (dynamic by time) and $\Delta C_{i,t}$ is the estimated economic loss during t in i-th subsystem.

$$\Delta C_{i,t} = |\mathrm{Px}_t \cdot \Delta E_{t,i}|. \tag{8.2}$$

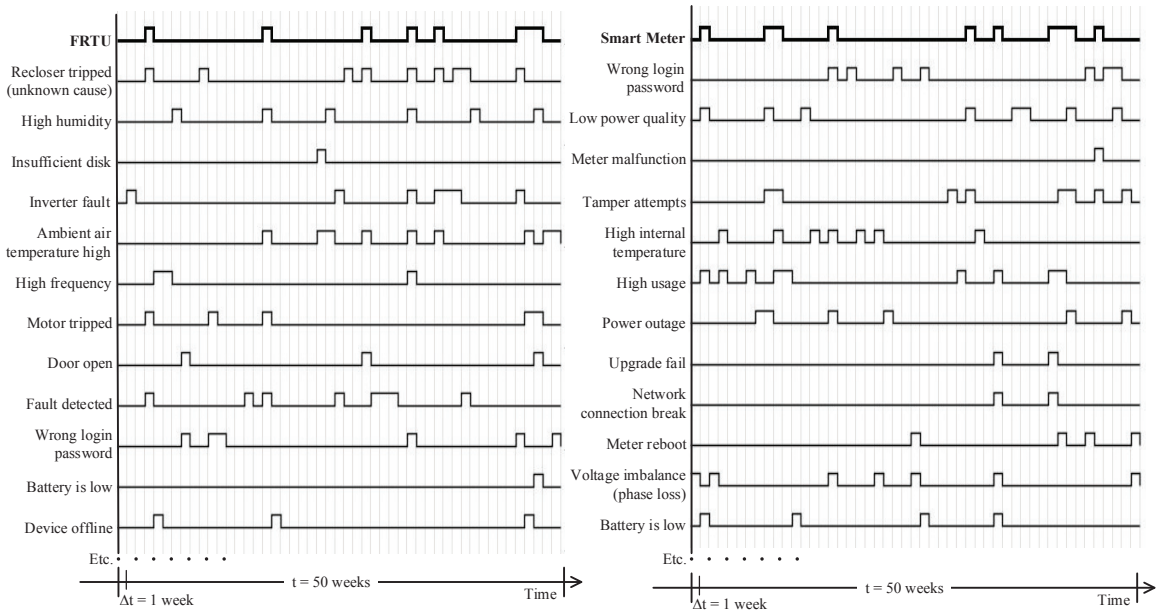

Figure 8.3: Binary representation of alarms and availability of FRTU and SM in time duration t.

8.3 Probabilistic Trust Model

Fig. 8.3 demonstrates an example of the binary representation of alarms and "heart-beat" of FRTU and SM in time duration t. Considering the availability and alarm

Table 8.1
Statistics of alarms for FRTU and SM in duration t.

FRTU			
Alarm Types	Threaten Levels	# of Times	Occurrence Prob.
Recloser Tripped	3	9	0.15
High Humidity	2	6	0.12
Insufficient Disk	1	1	0.01
Inverter Fault	2	7	0.14
High Temperature	2	9	0.20
High Frequency	2	3	0.06
Motor Tripped	3	5	0.10
Door Open	1	3	0.05
Fault Detected	3	8	0.15
Wrong Login Password	1	6	0.2
Low Battery	1	1	0.01
Device Offline	3	3	0.05
Smart Meter			
Wrong Login Password	1	7	0.15
Low Power Quality	2	8	0.15
Meter Malfunction	3	1	0.01
Tamper Attempts	3	8	0.15
High Temperature	2	7	0.13
High Usage	1	9	0.15
Power Outage	2	6	0.05
Upgrade Fail	3	2	0.15
Connection Break	3	2	0.02
Meter Reboot	2	4	0.03
Voltage Imbalance	2	7	0.05
Low Battery	1	4	0.01

results of a device as a binary event, we assume 0 represents "alive" or no alarms and 1 represents "dead" or alarm occurs. In this example, there are multiple types of alarms and t includes 50 Δts, which is the observation time units. Table 8.1 illustrates the corresponding statistical results of alarms for FRTU and SM as shown in Fig. 8.3. It should be noted that, in FRTU, the measurement will be recorded every 1 to 3 seconds while the recording is usually counted every 15-min in SMs, therefore, a time

normalization should be performed first to extend the recording time as 15-min either in the FRTU and then do the statistics for the whole observation time window. As mentioned before, alarms, which include the cyberattack and malware information, are the core factor to estimate the reliability of a device, the device trust model can be established according to alarm feedbacks. The mathematical model used to handle the dynamics of trust is based on probabilities transition. According to the statistic results of alarms in the archived records, the probability that an alarm occurs during a time duration t can be estimated as (8.3):

$$p_{\mathrm{A}} = 1 - \prod(1 - q_m),\ m = 1, 2, 3, \cdots . \tag{8.3}$$

where q_m represents the occurrence probability of type m alarm, which can be estimated from the statistics and observations of archived data. According to the values in the last column in Table 8.1, the p_{A} for the FRTU can be calculated as 0.738 and 0.618 for the SM.

Since there are multiple alarm properties with different priorities/severities in t, alarms are divided into 3 levels with different probabilities that may affect the device: 1) Level 1, almost no effect on normal operation, such as the wrong login password, the probability that the Level 1 alarm will affect the availability of the device is set as $\delta_{\mathrm{L}_1} = 0.01$; 2) Level 2, may affect the normal operation, such as high temperature,

the probability that these alarms will affect the availability of the device is set as $\delta_{\mathrm{L}_2} = 0.5$; and 3) Level 3, highly possible affect the normal operation, such as fault detected and the device unavailable probability is set as $\delta_{\mathrm{L}_3} = 0.95$. Therefore, the number of alarms in t, $N_{\mathrm{A},t} = N_{\mathrm{A},t,\mathrm{L}_1} + N_{\mathrm{A},t,\mathrm{L}_2} + N_{\mathrm{A},t,\mathrm{L}_3}$, where $N_{\mathrm{A},t}$ is the number of alarms occurred in t and $N_{\mathrm{A},t,\mathrm{L}_k}$ represents the number of alarms in level k. Then, the unavailable probability of the device in t according to different threaten levels' alarms can be represented by the conditional probability expression as (8.4):

$$\begin{aligned} p_{\mathrm{U},t} &= \sum^{k} p_t(\mathrm{U}|\mathrm{L}_k) \cdot p_t(\mathrm{L}_k) \\ &= \sum^{k} \delta_{t,\mathrm{L}_k} \cdot \frac{N_{\mathrm{A},t,\mathrm{L}_k}}{N_{\mathrm{A},t}},\ k = 1, 2, 3. \end{aligned} \tag{8.4}$$

In this example, the $p_{\mathrm{U},t}$ for the FRTU can be estimated as 0.596 and the value for the SM is 0.439. According to (8.3) and (8.4), the probabilistic trust model based on alarm responses in duration t can be represented as (8.5):

$$\tau_t = 1 - P_{\mathrm{A},t} \cdot P_{\mathrm{U},t}, \tag{8.5}$$

the trust model τ_t for j-th FRTU in t is represented as $\alpha_{t,j}$ and the alarm based trust model for l-th SM in t is $\beta_{t,l}$ in (8.1). The minimum observation time window for both FRTU and the SM is one week. In order to estimate the anomalous energy losses $\Delta E_{t,i}$ in (8.1), the recording time unit for the FRTU and SM should be normalized

first. In the example, $\alpha_t = 0.560$ and $\beta_t = 0.729$ for 50 weeks.

Following the increasing number of alarms in a device, the trust gradually decreases, on the contrary, if the number of alarms decreases in the next observation time window, the trust will gradually increase. Therefore, the accumulated probabilistic trust model should be affected by the trust value from the previous estimation duration and can be defined as (8.6):

$$\tau_a = \begin{cases} \tau_t \cdot \tau_{t-1}, \text{ if } N_{\mathrm{A},t} \geq N_{\mathrm{A},t-1}, \\ 1 - (1-\tau_t) \cdot (1-\tau_{t-1}), \text{ if } N_{\mathrm{A},t} < N_{\mathrm{A},t-1}, \end{cases} \tag{8.6}$$

$$t = 1, 2, 3 \cdots n.$$

The trust value of each FRTU and the SM should be regularly adjusted to reflect changes in the timely recording results. According to alarms statistics, the trust value from $\alpha_{t,j}$ and $\beta_{t,l}$ can be used to define four different cases.

† **Case 1:** There are no alarms from the primary network in a duration t, trust the measurement from the primary network but suspect the measurement from the secondary network, $\alpha_{t,j} = 1$, the $\beta_{t,l}$ is the decision variable;

† **Case 2:** There are no alarms from the secondary network in a duration t,

suspect the measurement from the primary network but trust the measurement from the secondary network, $\beta_{t,l} = 1$, the $\alpha_{t,j}$ is the decision variable;

† **Case 3:** If alarms received from both the primary and secondary network, suspect the measurements from the primary and the secondary network, both the $\alpha_{t,j}$ and $\beta_{t,l}$ are decision variables;

† **Case 4:** If the SCADA and AMI systems are unavailable, e.g. losing communication, system damaged, and/or unauthorized configuration change, $\alpha_{t,j} = 0, \beta_{t,l} = 0$, the subsystem can infer the consumption information from the allocation factor (AF) schemes.

Since Case 4 is a rare case and the system security logs and alerts can detect this anomaly, Cases 1 to 3 would be of interest to the massive tampering analysis and thus are the major concern of the proposed framework.

8.4 Case Study

As shown in Fig. 8.4 (a) is the tree representation of a 7-node topology. Two FRTUs are equipped ahead of nodes 2 and 7 in the primary network, three home SMs in the secondary network are connected on the lateral between these two FRTUs. The power injection direction is from the root node (node 1) to other downstream nodes.

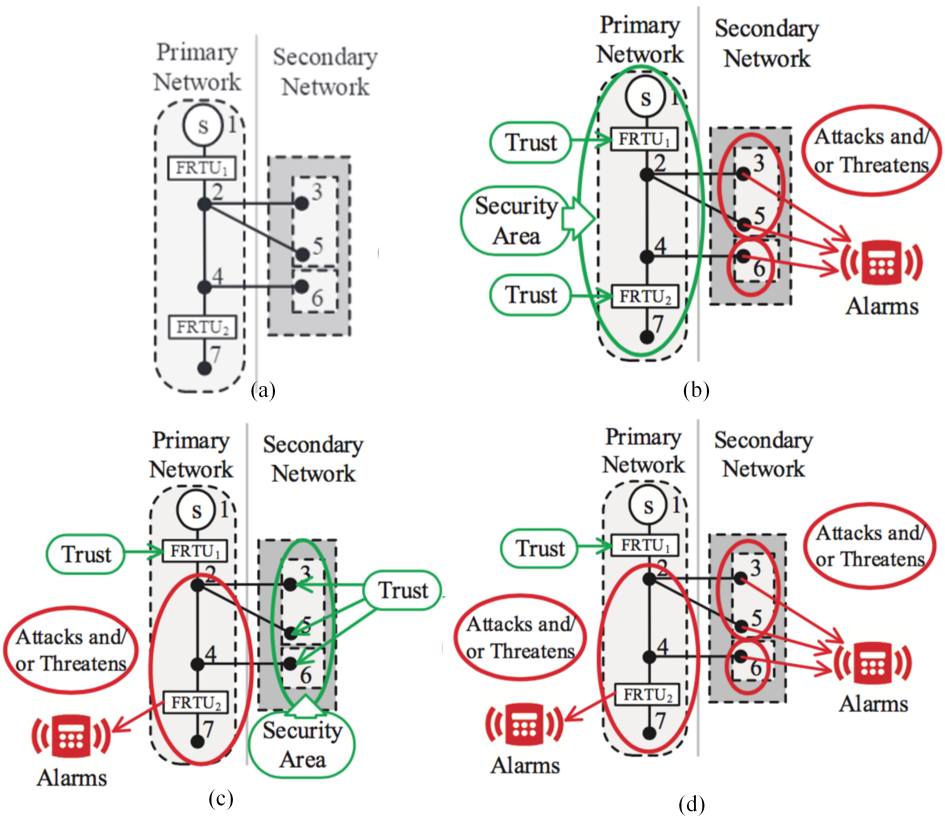

Figure 8.4: Spanning tree for a 7-node example topology with schematic diagrams for three cases.

The corresponding communication networks are also indicated in this representation. The following assumptions are made to facilitate the inference of massive tampering in this case:

1. All customers in the test subsystem are equipped with IP-based energy meters (smart meters);

2. The root FRTU/RTU is installed with complete security strategies to protect it from any cyberattack or malware tampering;

3. Detectable anomalies are from on-going attacks;

4. Most of the existing FRTUs and SMs use different communication networks, i.e., SCADA network versus AMI network, separated by frame relay switches, and different communication protocols (e.g., DNP3 versus ANSI C12.18). There is no direct communication link between the FRTUs and SMs;

5. There is no switching and system reconfiguration in the network while performing anomaly inference. Thus, network topology and parameters of both primary and secondary distribution networks are accurate in real time;

6. According to the statistic and observation results, the alarm occurrence probability q_m in Level 1 is 0.1, in Level 2 is 0.05, and 0.01 for Level 3;

7. The probability that the Level 1 alarm will affect the availability of the device is less or equal to 0.01, the probability for Level 2 alarm is around 0.5, and the Level 3 alarm has 95% possible to affect the device.

8.4.1 Case 1

An assumption, in this case, is the whole SCADA system has been developed with security strategies to protect it from the cyber attacks or any other threatens ($\alpha_j = 1$), while the validation for home SMs remains in the early stage. As shown in Fig. 8.4 (b), massive alarms occurred in HAN and/or NAN due to multiple types of attacks

or threatens that caused the $\Delta E_{t,i}$. Assume the unit observation time window is one month and the study duration is one season. The randomly generated array of alarms follows the Poisson distribution. The number of the different alarms (12 types) in the four observation time windows for SMs 3, 5, and 6 are illustrated in Fig. 8.5, separately. The curve figure demonstrates the monthly trust value for each observation window from (8.5) and the accumulated value according to (8.6) of these three SMs. In the estimation of accumulated trust value, the start point is assumed as the trust value generated from the first time window. The monthly trust values in each time window and the estimated accumulated trustworthy adjustment of case 1 are shown in Table 8.2.

8.4.2 Case 2

The security levels of SCADA and AMI system, in this case, are completely opposite than Case 1: the metering information from home SMs is validated because of the trustable communications in NAN/HAN ($\beta_{t,l} = 1$), but the WAN might experience attacks either from the physical or electronic layers as shown in Fig. 8.4 (c). The downstream FRTU ($FRTU_2$) received multiple alarms. The alarms types are different than that in SMs but also following the binary pattern. Fig. 8.6 demonstrates the statistic result of received alarms associated with the probabilistic trust evaluations in the j-th FRTU. The trustworthy adjustment of the affected FRTU is shown in

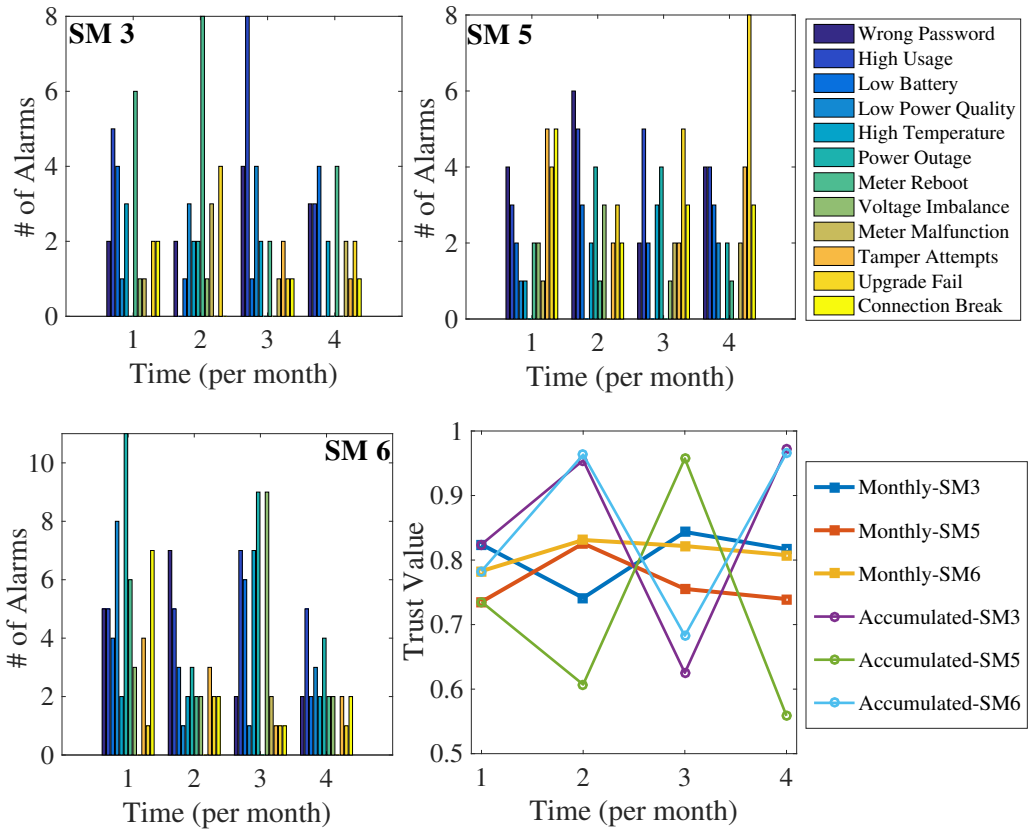

Figure 8.5: Statistic of alarms associated with the adjustment of trustworthy for each SM.

Table 8.2.

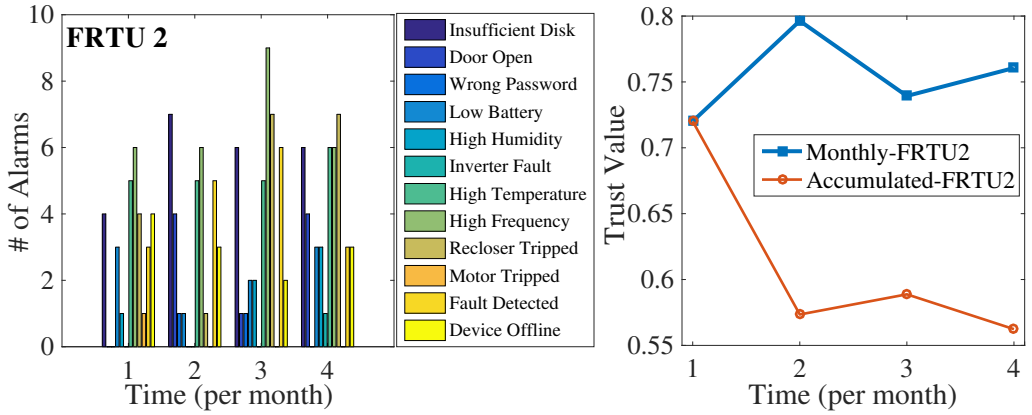

Figure 8.6: Statistic of alarms associated with the adjustment of trustworthy for the downstream FRTU.

8.4.3 Case 3

An undifferentiated cyberattack occurred in this subsystem, both the SCADA and the AMI networks are penetrated. As shown in Fig. 8.4 (d), the measurements from the downstream FRTU and SMs are all not trustable. Utilize the assumptions in Cases 1 and 2 to estimate the adjustments of trustworthy for each SM and the downstream FRTU. The randomly generated alarms also following the Poisson distribution and the trust variations in the observation time window for each SM and the downstream FRTU are illustrated in Fig. 8.7 and the numerical trust values of this case are recorded in Table 8.2.

Table 8.2 recorded the monthly and accumulated trust summaries for three cases. It can be seen that the trust value varies completely according to the number and the severity of alarms in the device (the heartbeat of the device) during a specific observation time period.

In the pre-defined data structure to perform the power flow analysis based on the example topology, the loads in the secondary network are defined as three-phase unbalanced with a wye connection type. The phase conductor of the overhead line in this example subsystem is selected as the *336,400 26/7 ACSR* while the neutral conductor is the *4/0 6/1 ACSR*. After defining the distance of each edge and the

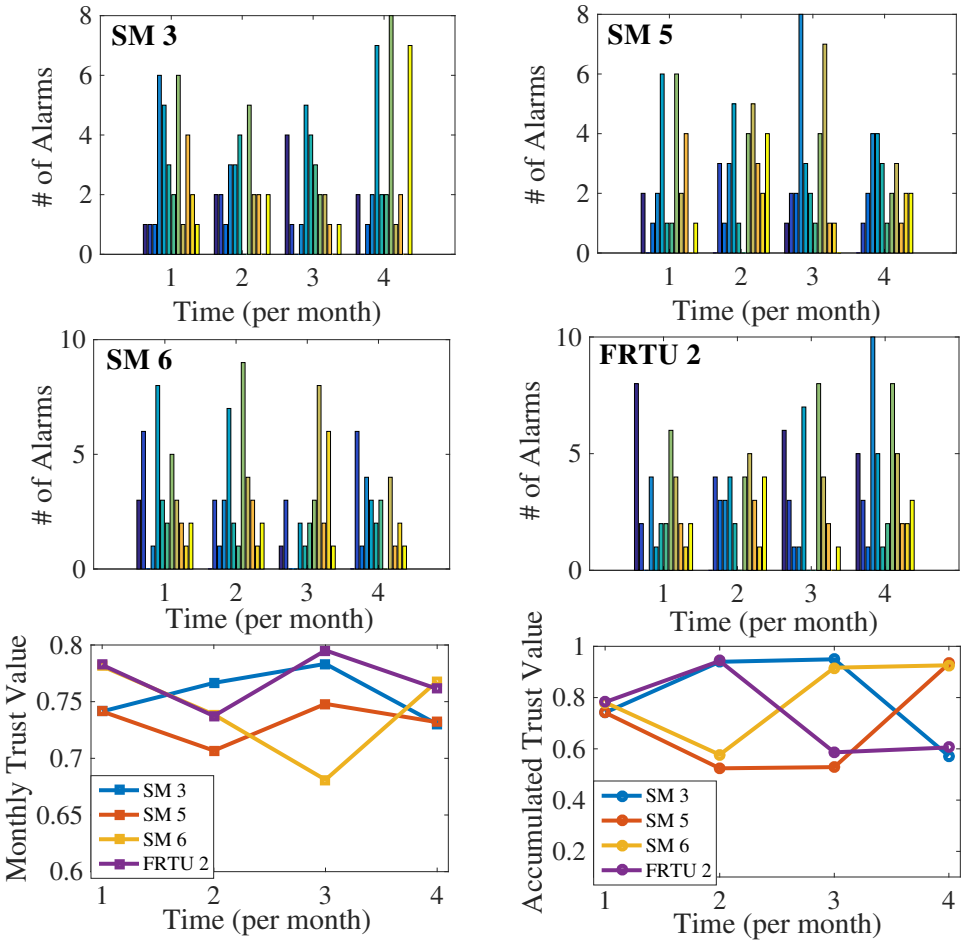

Figure 8.7: Statistic of alarms associated with the adjustment of trustworthy for each EM and the downstream FRTU.

Table 8.2
Monthly and accumulated trust summaries for three cases.

	Case 1			Case 2	Case 3			
	SM_3	SM_5	SM_6	$FRTU_2$	SM_3	SM_5	SM_6	$FRTU_2$
Monthly	0.824	0.735	0.783	0.720	0.741	0.741	0.781	0.782
	0.741	0.826	0.831	0.796	0.767	0.707	0.739	0.737
	0.844	0.755	0.822	0.739	0.783	0.748	0.681	0.795
	0.817	0.740	0.808	0.761	0.730	0.731	0.768	0.761
Accumulated	0.824	0.735	0.78	0.720	0.741	0.741	0.781	0.782
	0.955	0.607	0.963	0.574	0.940	0.524	0.577	0.943
	0.625	0.957	0.683	0.589	0.949	0.529	0.917	0.586
	0.971	0.559	0.966	0.562	0.572	0.933	0.926	0.605

source voltage, a graph-based power flow can calculate the voltages and currents of all nodes. Under the same conditions, input the topology and relative properties into another simulator (GridLAB-D) to generate the energy consumptions approximate to the real environment as the measurements. In the ideal case, no cyber attack occurred in this subsystem ($\alpha_j = 1, \beta_l = 1$), the estimated energy measurements and the average consumption from power flow results for each metering points in one month are demonstrated in row 1 of Table 8.3. The light yellow columns show the normal technical losses in power distribution.

Table 8.3
Estimated power measurements for each FRTU and EM in the case subsystem.

	Measured						Power Flow						
	SCADA		AMI			Loss	SCADA		AMI			Loss	%
	$FRTU_1$	$FRTU_2$	SM_3	SM_5	SM_6		$FRTU_1$	$FRTU_2$	SM_3	SM_5	SM_6		
Actual Value	989.79	208.47	261.72	269.83	247.54	2.23	985.06	217.79	254.62	270.12	239.46	3.0604	0.31%
	987.94	231.91	244.29	268.53	241.33	1.88	994.12	241.20	240.14	270.24	239.56	2.9790	0.30%
	985.67	233.66	242.10	268.21	237.97	3.73	986.39	232.85	240.77	270.15	239.49	3.1331	0.32%
	986.13	229.45	243.35	267.99	242.75	2.59	979.67	220.63	246.73	270.08	239.43	2.7963	0.29%
	Sensitivity						Adjusted Measurements						
	SCADA		AMI			Subsystem	SCADA		AMI			Var.	%
	$FRTU_1$	$FRTU_2$	SM_3	SM_5	SM_6		$FRTU_1$	$FRTU_2$	SM_3	SM_5	SM_6		
Case 1	100%	100%	82.4%	73.5%	78.3%	47.42%	989.79	208.47	215.66	198.33	193.82	171.28	17.30%
	100%	100%	74.1%	82.6%	83.1%	50.86%	987.94	231.91	181.02	221.81	200.55	150.77	15.26%
	100%	100%	84.4%	75.5%	82.2%	52.38%	985.67	233.66	204.33	202.50	195.61	145.84	14.80%
	100%	100%	81.7%	74.0%	80.8%	48.85%	986.13	229.45	198.82	198.31	196.14	160.82	16.31%
	Sensitivity						Adjusted Measurements						
	SCADA		AMI			Subsystem	SCADA		AMI			Var.	%
	$FRTU_1$	$FRTU_2$	SM_3	SM_5	SM_6		$FRTU_1$	$FRTU_2$	SM_3	SM_5	SM_6		
Case 2	100%	72.0%	100%	100%	100%	72.0%	989.79	150.10	261.72	269.83	247.54	58.37	5.90%
	100%	79.6%	100%	100%	100%	79.6%	987.94	184.60	244.29	268.53	241.33	47.31	4.79%
	100%	73.9%	100%	100%	100%	73.9%	985.67	172.67	242.10	268.21	237.97	60.99	6.19%
	100%	76.1%	100%	100%	100%	76.1%	986.13	174.61	243.35	267.99	242.75	54.84	5.56%
	Sensitivity						Adjusted Measurements						
	SCADA		AMI			Subsystem	SCADA		AMI			Var.	%
	$FRTU_1$	$FRTU_2$	SM_3	SM_5	SM_6		$FRTU_1$	$FRTU_2$	SM_3	SM_5	SM_6		
Case 3	100%	74.1%	74.1%	78.1%	78.2%	33.53%	989.79	163.02	193.93	199.94	193.33	237.34	23.98%
	100%	76.7%	70.7%	73.9%	73.7%	29.53%	987.94	170.92	187.37	189.85	178.34	259.58	26.27%
	100%	78.3%	74.8%	68.1%	79.5%	31.71%	985.67	185.76	189.56	200.62	162.06	243.94	24.75%
	100%	73.0%	73.1%	76.8%	76.1%	31.19%	986.13	174.61	177.65	195.90	186.43	248.95	25.25%

⋆ Unit of measurements is kWh; Var. represents variation.

Rows 2 to 4 demonstrate the sensitivities and adjusted measurements for three cases. The trust percentages illustrated with green columns in sensitivity part of each case are from Table 8.2. The calculated adjusted measurements are combining the trust values and the measured information from row 1. The trustworthy parameter of the subsystem is following (8.7)

$$1 - \prod (1 - \tau_j), \tag{8.7}$$

where τ_j is the trust value of the j-th device in the subsystem. In Case 1, the varied trust values of these three SMs lead to adjusted values in energy consumptions. The total energy variation of the subsystem excludes the normal energy loss from the measurements. In Case 2, the possible energy consumption in the downstream FRTU is changed and both the SCADA and the AMI system are not trustable in Case 3. The actual measurement variations must less than the calculated variations. On the contrary, once the variation in a subsystem exceeds the potential variation value, there must exist massive tampering (more than 20% variation).

Chapter 9

Conclusion

One of the critical milestones on smart grid vision is to deploy sensing devices within an electrical network. As evident by today's mobile phone technologies, the traffic information can be inferred from the movement of mobile users to determine their whereabouts. The proposed agent-based framework in Chapter 2 represents the observable dynamics of consumers' energy consumption and movements which are used to correlate the patterns of two interdependencies. The adaptive parameterization of occupancy-consumption aggregation has demonstrated the feasibility of the proposed correlation framework. The finite mixtures of regression models are adjusted by correlating with data available from AMI, on-site meter reading, SCADA measurements or historical datasets.

The smart meter deployment was aimed to deploy on every household. The increased metering points within a distribution network can improve observability. However, it would involve an initial investment and ongoing maintenance costs. The heterogeneous framework can be utilized to enhance load models and observability that incorporate the dynamics of occupants' movements and how the existence patterns may influence the energy consumptions. As each load may have unique patterns that can be inferred from their activities, the proposed OP pairs can set as a statistical reference for inferring consumption usages. The load profile enrichment can be established through the infusion of historical energy consumption datasets.

The proposed heterogeneous framework can be implemented in multiple ways. One way is to enable a data exchange path between the communication providers and utilities. The other alternative is to promote mobile app by utilities that would engage and incentivize their customers. Some utilities may have metering data from their distribution SCADA system. There are several metering points within a feeder that would provide the data acquisition. Utilities might have to invest additional network infrastructure and security measures to ensure the confidentiality of real-time occupant datasets. Privacy is important and this issue can be mitigated. The identity of occupants at the certain location will be masked at the level of service providers that do not include anything but their physical location at the time. To strengthen the privacy of household consumers, the utility company can strictly enforce a non-disclosure agreement with employees.

The correlation framework can be utilized to estimate an unmetered load from other similar profiles. Such inference provides a means to establish new load profiles that can create a new application in the distribution control center. This can eliminate on-site data collection visits as a result of labor savings. This preliminary work also offers to improve system observability with other interdependent networks. The extension of other load types, such as industrial or commercial, can be enhanced by relating with thousands of metered and unmetered loads within a feeder.

In Chapter 3, it proposed an anomaly detection and localization technique using switching procedures based on the graph theory. The profile-based anomaly detection method is utilized to compare the consumption value displayed on the branch head with the summation of all meters readings in its downstream. After localizing the abnormal load node, the comparison of consumption pattern for all meters connected with this node will be performed to find the tampered meter.

In the process of localizing the abnormal load node, the distribution network is converted to a spanning tree to demonstrate the connection states in the topology and to avoid generating loops during switching procedures, and then convert the spanning tree to incidence or adjacency matrix for future analysis. An anomaly localization algorithm is introduced to find the tampered load node. In addition, the cost consideration is mentioned to balance the benefit of localizing the tampered meter. In this work, a virtual distribution network is applied in the case study, a real distribution

network will be used to validate the algorithm in the future. Since the main content of this work is in the strategy of switching the openable edges while meeting electrical and operational constraints, the search process could be an exponential time search or only spend a linear time. The corresponding combination strategy and search method will be illustrated in a future analysis. Some simulation software such as the geographic information system (GIS) can be utilized to construct the real topology. Another vector to store the open switch information instead of changing the input incidence matrix during each iteration will be added. An agent-based model to generate random tampered nodes and random anomaly frequencies in the provided network will be conducted. This model can also generate the consumption dataset for estimating the effect of anomalies in the topology and calculating the cost due to the fraudulent activities. The anomaly localization algorithm will be improved to adapt the situations that more than one tampering exists in the same feeder or in the whole network. Furthermore, a heuristic algorithm developed to replace the DFS method in the original work to improve the search efficiency in a spanning tree.

Massive tampering of electrical metering is a notorious problem in the electric power system, which causes great economic losses and threatens the reliability of the power system. In smart grids, smart meters and even the FRTU/RTU with a higher-level security level may potentially be attacked or tampered to cause certain anomalous losses. It is challenging to identify large-scale malicious meters when there are a large number of users. Finding efficient measurements for detecting falsified consumption

datasets and identifying the tampered load(s) have been active researchers in recent years. Statistical studies have shown that malware is one of the main reasons leading to massive tampering and even threatening the entire system.

In addition to providing real-time data logging, modern grid metering equipment can provide anomalous feedback. The SCADA system in the primary network realizes data transmission through the WAN network, while the AMI system in the secondary network communicates through HAN and NAN. The two are independent networks, and any system or measurement device in the power grid encounters network attacks or any threat from other physical or electronic layers. It will send alarms to the control center and archive the relevant event logs. This paper creates a probabilistic trust model to adjust and estimate the deviation of measured values in the system or a subsystem within an observation time window by extracting the statistics of alarms and related event logs. Thus realizing the inference of whether there is massive tampering. At this stage, only one inference result is proposed in this paper. It does not propose prevention, defense, and/or modification measures. In the future study, through the reconfiguration of the system physical or network topology, combining with the trust model and adjusted measurements at each node, the tampered lumped load could be localized. Furthermore, the corresponding maintenance and defense strategies should be provided.

References

[1] Z. Shuai, C. Shen, X. Yin, X. Liu, and Z. J. Shen, "Fault analysis of inverter-interfaced distributed generators with different control schemes," *IEEE Trans. Power Del.*, vol. 33, no. 3, pp. 1223–1235, Jun. 2018.

[2] R. C. A. Palacio, M. Mezaroba, and J. R. Pinheiro, "VSG based control application for inverter-interfaced distributed generators in microgrids," in *Brazilian Power Electronics Conf. (COBEP)*, Nov. 2017, pp. 1–6.

[3] Z. Zeng, H. Yang, S. Tang, and R. Zhao, "Objective-oriented power quality compensation of multifunctional grid-tied inverters and its application in microgrids," *IEEE Trans. Power Electron.*, vol. 30, no. 3, pp. 1255–1265, Mar. 2015.

[4] Z. Zeng, R. Zhao, and H. Yang, "Coordinated control of multi-functional grid-tied inverters using conductance and susceptance limitation," *IET Power Electronics*, vol. 7, no. 7, pp. 1821–1831, Jul. 2014.

[5] Z. Zeng, X. Li, and W. Shao, "Multi-functional grid-connected inverter: upgrading distributed generator with ancillary services," *IET Renewable Power Generation*, vol. 12, no. 7, pp. 797–805, 2018.

[6] S. Alyami, C. Wang, and C. Fu, "Development of autonomous schedules of controllable loads for cost reduction and pv accommodation in residential distribution networks," in *Electrical Power and Energy Conf. (EPEC)*, Oct. 2015, pp. 81–86.

[7] T. Shintai, Y. Miura, and T. Ise, "Oscillation damping of a distributed generator using a virtual synchronous generator," *IEEE Trans. Power Del.*, vol. 29, no. 2, pp. 668–676, Apr. 2014.

[8] J. Alipoor, Y. Miura, and T. Ise, "Power system stabilization using virtual synchronous generator with alternating moment of inertia," *IEEE Journal of Emerging and Selected Topics in Power Electronics*, vol. 3, no. 2, pp. 451–458, Jun. 2015.

[9] K. P. Schneider, B. A. Mather, B. C. Pal, C. W. Ten, G. J. Shirek, H. Zhu, J. C. Fuller, J. L. R. Pereira, L. F. Ochoa, L. R. de Araujo, R. C. Dugan, S. Matthias, S. Paudyal, T. E. McDermott, and W. Kersting, "Analytic considerations and design basis for the ieee distribution test feeders," *IEEE Trans. Power Syst.*, vol. 33, no. 3, pp. 3181–3188, May 2018.

[10] G. W. Chang, S. Y. Chu, and H. L. Wang, "An improved backward/forward

sweep load flow algorithm for radial distribution systems," *IEEE Trans. Power Syst.*, vol. 22, no. 2, pp. 882–884, May 2007.

[11] Y. Zhu and K. Tomsovic, "Adaptive power flow method for distribution systems with dispersed generation," *IEEE Trans. Power Del.*, vol. 17, no. 3, pp. 822–827, Jul. 2002.

[12] S. Ghosh and D. Das, "Method for load-flow solution of radial distribution networks," *IEE Proceedings - Generation, Transmission and Distribution*, vol. 146, no. 6, pp. 641–648, Nov. 1999.

[13] D. Shirmohammadi, H. W. Hong, A. Semlyen, and G. X. Luo, "A compensation-based power flow method for weakly meshed distribution and transmission networks," *IEEE Trans. Power Syst.*, vol. 3, no. 2, pp. 753–762, May 1988.

[14] W. H. Kersting, *Distribution System Modeling and Analysis*, 3rd ed. Taylor & Francis Group., 2011.

[15] M. Abrar, M. A. Tahir, R. Masroor, and H. M. U. Hamid, "Real time smart grid load management by integrated and secured communication," in *Intl. Conf. on Innovative Trends in Computer Engineering (ITCE)*, Feb. 2018, pp. 253–257.

[16] A. Qaddus and A. A. Minhas, "Wireless communication a sustainable solution for future smart grid networks," in *Intl. Conf. on Open Source Systems Technologies (ICOSST)*, Dec. 2016, pp. 13–17.

[17] J. A. Dias, P. J. A. Serni, and E. P. Godoy, "Study of communication between distributed generation devices in an smart grid environment," *IEEE Latin America Transactions*, vol. 16, no. 3, pp. 777–784, Mar. 2018.

[18] V. Kouhdaragh, I. S. Amiri, and S. Seyedi, "Smart grid load balancing methods to make an efficient heterogeneous network by using the communication cost function," *IET Networks*, vol. 7, no. 3, pp. 95–102, 2018.

[19] U. Ahsan and A. Bais, "Distributed smart home architecture for data handling in smart grid," *Canadian Journal of Electrical and Computer Engineering*, vol. 41, no. 1, pp. 17–27, 2018.

[20] B. Nasiri, C. Wagner, U. Hger, and C. Rehtanz, "Distribution grid planning considering smart grid technologies," *CIRED - Open Access Proceedings Journal*, vol. 2017, no. 1, pp. 2228–2232, 2017.

[21] R. Kappagantu and S. A. Daniel, "Challenges and issues of smart grid implementation: A case of indian scenario," *Journal of Electrical Systems and Information Technology*, vol. 5, no. 1, pp. 1–15, Feb. 2018.

[22] R. Ma, H.-H. Chen, Y.-R. Huang, and W. Meng, "Smart grid communication: Its challenges and opportunities," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 36–46, Mar. 2013.

[23] E. Bou-Harb, C. Fachkha, M. Pourzandi, M. Debbabi, and C. Assi, "Communication security for smart grid distribution networks," *IEEE Communications Magazine*, vol. 51, no. 1, pp. 42–49, Jan. 2013.

[24] C. H. Lin, S. J. Chen, C. L. Kuo, and J. L. Chen, "Non-cooperative game model applied to an advanced metering infrastructure for non-technical loss screening in micro-distribution systems," *IEEE Trans. Smart Grid*, vol. 5, no. 5, pp. 2468–2469, Sept. 2014.

[25] T. S. Zhan, S. J. Chen, C. C. Kao, C. L. Kuo, J. L. Chen, and C. H. Lin, "Non-technical loss and power blackout detection under advanced metering infrastructure using a cooperative game based inference mechanism," *IET Generation, Transmission Distribution*, vol. 10, no. 4, pp. 873–882, Mar. 2016.

[26] R. Blom, K. Norrman, M. Nslund, S. Rommer, and B. Sahlin. (2010) Security in the evolved packet system. [Online]. Available: https://www.eit.lth.se/fileadmin/eit/courses/eitn50/Literature/security_eps.pdf.

[27] J. Naruchitparames, M. H. Gunes, and C. Y. Evrenosoglu, "Secure communications in the smart grid," in *Proc. Consumer Communications and Networking Conf.*, Jan. 2011, pp. 1171–1175.

[28] F. Knirsch, D. Engel, M. Frincu, and V. Prasanna, "Model-based assessment for balancing privacy requirements and operational capabilities in the smart

grid," in *IEEE Power Energy Society Innovative Smart Grid Technologies Conf. (ISGT)*, Feb. 2015, pp. 1–5.

[29] J. Giraldo, A. Crdenas, E. Mojica-Nava, N. Quijano, and R. Dong, "Delay and sampling independence of a consensus algorithm and its application to smart grid privacy," in *53rd IEEE Conf. on Decision and Control*, Dec. 2014, pp. 1389–1394.

[30] C. Neureiter, G. Eibl, A. Veichtlbauer, and D. Engel, "Towards a framework for engineering smart-grid-specific privacy requirements," in *39th Annual Conf. of the IEEE Industrial Electronics Society*, Nov. 2013, pp. 4803–4808.

[31] C. Jouvray, G. Pellischek, and M. Tiguercha, "Impact of a smart grid to the electric vehicle ecosystem from a privacy and security perspective," in *World Electric Vehicle Symposium and Exhibition (EVS27)*, Nov. 2013, pp. 1–10.

[32] R. Kumar and D. Saxena, "Fault analysis of a distribution system embedded with plug-in electric vehicles," in *Recent Developments in Control, Automation Power Engineering (RDCAPE)*, Oct. 2017, pp. 230–234.

[33] T. F. Moraes, L. Lovisolo, and L. F. C. Monteiro, "Fault location in distribution systems from analysis of the energy of sequence component waveforms," *IET Generation, Transmission Distribution*, vol. 12, no. 9, pp. 1951–1960, 2018.

[34] A. Ali, A. Q. Khan, B. Hussain, M. T. Raza, and M. Arif, "Fault modelling

and detection in power generation, transmission and distribution systems," *IET Generation, Transmission Distribution*, vol. 9, no. 16, pp. 2782–2791, 2015.

[35] N. Cho, R. Bhat, and J. H. Lee, "Smart fault isolation and service restoration under consideration of abnormal conditions in distribution systems," in *Intl. Symposium on Smart Electric Distribution Systems and Technologies (EDST)*, Sept. 2015, pp. 362–367.

[36] K. Jia, T. Bi, B. Liu, E. Christopher, D. W. P. Thomas, and M. Sumner, "Marine power distribution system fault location using a portable injection unit," *IEEE Trans. Power Del.*, vol. 30, no. 2, pp. 818–826, Apr. 2015.

[37] M. K. Yoon, S. Mohan, J. Choi, M. Christodorescu, and L. Sha, "Learning execution contexts from system call distribution for anomaly detection in smart embedded system," in *IEEE/ACM Second Intl. Conf. on Internet-of-Things Design and Implementation (IoTDI)*, Apr. 2017, pp. 191–196.

[38] M. Gui, A. Pahwa, and S. Das, "Anomaly detection in animal-related failures in overhead distribution systems," in *39th North American Power Symposium*, Sept. 2007, pp. 498–504.

[39] T. Akhtar, B. B. Gupta, and S. Yamaguchi, "Malware propagation effects on SCADA system and smart power grid," in *IEEE Intl. Conf. on Consumer Electronics (ICCE)*, Jan. 2018, pp. 1–6.

[40] H. Hilal and A. Nangim, "Network security analysis SCADA system automation

on industrial process," in *Intl. Conf. on Broadband Communication, Wireless Sensors and Powering (BCWSP)*, Nov. 2017, pp. 1–6.

[41] E. Ciancamerla, M. Minichino, and S. Palmieri, "Modelling SCADA and corporate network of a medium voltage power grid under cyber attacks," in *Intl. Conf. on Security and Cryptography (SECRYPT)*, Jul. 2013, pp. 1–12.

[42] E. Byres, A. Ginter, and J. Langill. (2010) White Paper: How stuxnet spreads-a study of infection paths in best practice systems. Tofino Security. [Online]. Available: https://www.slideshare.net/YuryChemerkin/how-stuxnet-spreads-a-study-of-infection-paths-in-best-practice-systems.

[43] A. A. Falaye, O. Osho, M. I. Emehian, and S. Ale, "Dynamics of SCADA system malware: Impacts on smart grid electricity networks and countermeasures," in *Proc. Intl. Conf. on Information and Communication Technology and Its Applications*, Nov. 2016, pp. 139–145.

[44] S. K. Singh, R. Bose, and A. Joshi, "Energy theft detection in advanced metering infrastructure," in *IEEE 4th World Forum on Internet of Things (WF-IoT)*, Feb. 2018, pp. 529–534.

[45] A. S. Metering, S. Visalatchi, and K. K. Sandeep, "Smart energy metering and power theft control using arduino gsm," in *2nd Intl. Conf. for Convergence in Technology (I2CT)*, Apr. 2017, pp. 858–961.

[46] J. Smith. (2015) Smart meters take bite out of electricity theft. National

Geographic. [Online]. Available: http://news.nationalgeographic.com/news/energy/2011/09/110913-smart-meters-for-electricity-theft/.

[47] (2013) Using analytics to crack down on electricity theft. Deloitte. [Online]. Available: http://deloitte.wsj.com/cio/2013/12/02/using-analytics-to-crack-down-on-electricity-theft/.

[48] (2016, Dec.) Assessment of demand response and advanced metering. Federal Energy Regulatory Commission. [Online]. Available: https://www.ferc.gov/legal/staff-reports/2016/DR-AM-Report2016.pdf.

[49] (2016, Oct.) Form EIA-861. Energy Information Administration (EIA). [Online]. Available: https://www.eia.gov/electricity/data/eia861/index.html.

[50] (2014, Feb.) Advanced metering infrastructure and customer systems. Recovery Act Smart Grid Programs. [Online]. Available: http://www.smartgrid.gov/recovery_act/deployment_status/ami_and_customer_systems.

[51] X. Gu, H. cheng Wang, and J. Chen, "Application of rough set-based distribution network fault location approach in trouble call management system," in *China Intl. Conf. on Electricity Distribution*, Sept. 2012, pp. 1–5.

[52] C. S. Chang and F. S. Wen, "Tabu search based approach to trouble call analysis in LV power distribution," *IEE Proceedings - Generation, Transmission and Distribution*, vol. 145, no. 6, pp. 731–738, Nov. 1998.

[53] M. T. Tsay, W. M. Lin, and A. J. Hwang, "A reliability model based trouble call analysis involving old secondary circuits," in *Intl. Conf. on Power System Technology*, vol. 1, Aug. 1998, pp. 270–274.

[54] X. Zhang, W. Jouini, P. Leray, and J. Palicot, "Temperature-power consumption relationship and hot-spot migration for fpga-based system," in *IEEE/ACM Int'l Conf. on Cyber, Physical and Social Computing (CPSCom)*, Dec. 2010, pp. 392–397.

[55] Y. Saika and M. Nakagawa, "Dynamics of predicting temperature-humidity index and power consumption in small-scale system utilizing bayesian inference via the eap estimation," in *17th Intl. Conf. on Control, Automation and Systems (ICCAS)*, Oct. 2017, pp. 975–980.

[56] Y. Tarutani, K. Hashimoto, G. Hasegawa, Y. Nakamura, T. Tamura, K. Matsudax, and M. Matsuoka, "Reducing power consumption in data center by predicting temperature distribution and air conditioner efficiency with machine learning," in *IEEE Intl. Conf. on Cloud Engineering (IC2E)*, Apr. 2016, pp. 226–227.

[57] Y. Tarutani, K. Hashimoto, G. Hasegawa, Y. Nakamura, T. Tamura, K. Matsuda, and M. Matsuoka, "Temperature distribution prediction in data centers for decreasing power consumption by machine learning," in *IEEE 7th Intl. Conf.*

on Cloud Computing Technology and Science (CloudCom), Nov. 2015, pp. 635–642.

[58] V. Lakshmanan, M. Marinelli, A. M. Kosek, F. Sossan, and P. Nrgrd, "Domestic refrigerators temperature prediction strategy for the evaluation of the expected power consumption," in *IEEE PES ISGT Europe*, Oct. 2013, pp. 1–5.

[59] K. Yuuki, T. Sakano, and Y. Yokomizo, "Low power consumption solid state humidity sensor," *IEEE Trans.on Consumer Electronics*, vol. CE-29, no. 3, pp. 305–309, Aug. 1983.

[60] T. Din and S. Hillmansen, "Energy consumption and carbon dioxide emissions analysis for a concept design of a hydrogen hybrid railway vehicle," *IET Electrical Systems in Transportation*, vol. 8, no. 2, pp. 112–121, May 2018.

[61] (1999) Novelty Detection using Extreme Value Statistics. Stephen J. Roberts. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.1338&rep=rep1&type=pdf.

[62] (2002) Poster: An Extreme Value Theory Approach to Anomaly Detection (EVT-AD). Sandra G. Dykes. [Online]. Available: http://www.ieee-security.org/TC/SP2012/posters/An%20Extreme%20Value%20Theory%20Approach.pdf.

[63] M. Z. Z. Abiden and S. Z. Z. Abidin, "Introducing residual errors in accuracy

assessment for remotely sensed change detection," in *5th Intl. Colloquium on Signal Processing Its Applications*, Mar. 2009, pp. 89–92.

[64] (2018) The Residual Correction Method. Mathonline. [Online]. Available: http://mathonline.wikidot.com/the-residual-correction-method.

[65] C. A. Fantin, M. R. C. Castillo, B. E. B. de Carvalho, and J. B. A. London, "Using pseudo and virtual measurements in distribution system state estimation," in *IEEE PES Transmission Distribution Conf. and Exposition - Latin America (PES TD-LA)*, Sept. 2014, pp. 1–6.

[66] B. Hannaford, "Opportunities and problems in the electric distribution system," *Journal of the A.I.E.E.*, vol. 45, no. 2, pp. 180–184, Feb. 1926.

[67] G. M. Burt, J. R. McDonald, A. G. King, J. Spiller, D. Brooke, and R. Samwell, "Intelligent on-line decision support for distribution system control and operation," *IEEE Transactions on Power Systems*, vol. 10, no. 4, pp. 1820–1827, Nov. 1995.

[68] Y. Mao and K. Miu, "Switch placement to improve system reliability for radial distribution systems with distributed generation," in *IEEE Power Engineering Society General Meeting*, Jun. 2004, p. 890.

[69] M. I. Abouheaf, W. J. Lee, and F. L. Lewis, "Dynamic formulation and approximation methods to solve economic dispatch problems," *IET Generation, Transmission Distribution*, vol. 7, no. 8, pp. 866–873, Aug. 2013.

[70] S. M. Dean, "The design and operation of a metropolitan electrical system from the viewpoint of possible major shutdowns," *Electrical Engineering*, vol. 59, no. 10, pp. 575–579, Oct. 1940.

[71] F. S. Yao, X. Q. Zhang, Y. Zhang, and T. H. Wang, "Computer decision-making support system for power distribution network planning based on geographical information system," in *China Intl. Conf. on Electricity Distribution*, Dec. 2008, pp. 1–6.

[72] M. F. Alhajri and M. E. El-Hawary, "Exploiting the radial distribution structure in developing a fast and flexible radial power flow for unbalanced three-phase networks," *IEEE Transactions on Power Delivery*, vol. 25, no. 1, pp. 378–389, Jan. 2010.

[73] R. Stoicescu, K. Miu, C. O. Nwankpa, D. Niebur, and X. Yang, "Three-phase converter models for unbalanced radial power flow studies," *IEEE Transactions on Power Systems*, vol. 17, no. 4, pp. 1016–1021, Nov. 2002.

[74] C. W. Ten and Y. Tang, *Distribution Emergency Operation*. CRC Press Taylor & Francis Group, Aug. 2018, ch. 5.

[75] W. Sadiq and M. E. Orlowska, "Analyzing process models using graph reduction techniques," *Information Systems*, vol. 25, no. 2, pp. 117–134, Apr. 2000.

[76] H. Lin, Z. Zhao, H. Li, and Z. Chen, "A novel graph reduction algorithm

to identify structural conflicts," in *35th Annual Hawaii Intl. Conf. on System Sciences*, Jan. 2002, pp. 1–10.

[77] W. Sadiq and M. E. Orlowska, "Applying graph reduction techniques for identifying structural conflicts in process models," in *11th Intl. Conf. on CAiSE*, Jun. 1999, pp. 195–209.

[78] B. Venkatesh and R. Ranjan, "Data structure for radial distribution system load flow analysis," *IEE Proceedings - Generation, Transmission and Distribution*, vol. 150, no. 1, pp. 101–106, Jan. 2003.

[79] R. Parasher, "Load flow analysis of radial distribution network using linear data structure," *CoRR*, vol. abs/1403.4702, 2014.

[80] (2013, Oct.) Assessment of demand response and advanced metering staff report. Federal Energy Regulatory Commission (FERC). [Online]. Available: http://www.ferc.gov/legal/staff-reports/2013/oct-demand-response.pdf.

[81] (2012, Nov.) Smart meter deployments continue to rise. U.S. Energy Information Administration (EIA). [Online]. Available: http://www.eia.gov/todayinenergy/detail.cfm?id=8590.

[82] (2012, Dec.) Assessment of demand response and advanced metering staff report. Federal Energy Regulatory Commission (FERC). [Online]. Available: http://www.ferc.gov/legal/staff-reports/12-20-12-demand-response.pdf.

[83] Y. Tang, C.-W. Ten, C. Wang, and G. Parker, "Extraction of energy information from analog meters using image processing," *IEEE Trans. Smart Grid*, vol. 6, no. 4, pp. 2032–2040, Jan. 2015.

[84] L. Alexandera, S. Jiang, M. Murga, and M. C. Gonzalez, "Origindestination trips by purpose and time of day inferred from mobile phone data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 240–250, Sept. 2015.

[85] C. Chen, L. Bian, and J. Ma, "From traces to trajectories: How well can we guess activity locations from mobile phone traces?" *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 326–337, Sept. 2014.

[86] P. Widhalm, Y. Yang, M. Ulm, S. Athavale, and M. Gonzalez, "Discovering urban activity patterns in cell phone data," *Transportation Research Part C: Emerging Technologies*, vol. 42, no. 4, pp. 597–623, Jul. 2015.

[87] (2013, Apr.) Creating a module. U.S. Department of Energy (DOE) at Pacific Northwest National Laboratory (PNNL). [Online]. Available: http://gridlab-d.shoutwiki.com/wiki/Creating_a_module.

[88] T. Ryan, *Modern Regression Methods*, 2nd ed. John Wiley & Sons, Inc., 2009.

[89] (2016) Linear regression analysis of energy consumption data. Bizee Software. [Online]. Available: http://www.degreedays.net/regression-analysis.

[90] M. Blackwell. (2008, Dec.) Multiple hypothesis testing: The f-test. [Online]. Available: http://www.mattblackwell.org/files/teaching/ftests.pdf.

[91] A. Arif, Z. Wang, J. Wang, B. Mather, H. Bashualdo, and D. Zhao, "Load modeling - a review," *IEEE Trans. Smart Grid*, vol. PP, no. 99, pp. 1–15, May 2017.

[92] K. P. Schneider, J. C. Fuller, and D. P. Chassin, "Multi-state load models for distribution system analysis," *IEEE Trans. Power Syst.*, vol. 26, no. 4, pp. 2425–2433, Nov. 2011.

[93] Z. Gou, Z. J. Wang, and A. Kashani, "Home appliance load modeling from aggregated smart meter data," *IEEE Trans. Power Syst.*, vol. 30, no. 1, pp. 254–262, Jan. 2015.

[94] L. Chuan and A. Ukil, "Modeling and validation of electrical load profiling in residential buildings in singapore," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2800–2809, Sept. 2015.

[95] D. T. Nguyen, "Modeling load unvertainty in distribution network monitoring," *IEEE Trans. Power Syst.*, vol. 30, no. 5, pp. 2321–2328, Sept. 2015.

[96] S. Zhao and K. Zhang, "Observing individual dynamic choices of activity chains from location-based crowdsourced data," in *Proc. 95th Annual Conf. on Transportation Research Board*, Jan. 2016, pp. 1–20.

[97] T. Carmenate, M. Rahman, D. Leante, L. Bobadilla, and A. Mostafavi, "Modeling and analyzing occupant behaviors in building energy analysis using an information space approach," in *Proc. IEEE Intl. Conf. on Automation Science and Engineering (CASE)*, Aug. 2015, pp. 425–431.

[98] M. S. Gul and S. Patidar, "Understanding the energy consumption and occupancy of a multi-purpose academic building," *Energy and Buildings*, vol. 87, no. 1, pp. 155–165, Jan. 2015.

[99] R. Gulbinas, A. Khosrowpour, and J. Taylor, "Segmentation and classification of commercial building occupants by energy-use efficiency and predictability," *IEEE Trans. Smart Grid*, vol. 6, no. 3, pp. 1414–1424, May 2015.

[100] A. Lotfi, L. Jalil, and A. Al-Habaibeh, "Investigating occupant behaviour to improve energy efficiency in social housing: Case study," in *Proc. 9th Intl. Conf. on Intelligent Environments (IE)*, Jul. 2013, pp. 124–128.

[101] N. A. Zanjani, G. Lilis, G. Conus, and M. Kayal, "Energy book for buildings: Occupants incorporation in energy efficiency of buildings," in *Proc. Intl. Conf. on Smart Cities and Green ICT Systems (SMARTGREENS)*, May 2015, pp. 1–6.

[102] Y. Sakakura, "Household power consumption simulator with compact representation of occupant behaviors," in *Proc. IEEE Intl. Conf. on Smart Grid Communication*, Jan. 2015, pp. 170–175.

[103] B. J. Johnson, M. R. Starke, O. A. Abdelaziz, R. K. Jackson, and L. M. Tolbert, "A method for modeling household occupant behavior to simulate residential energy consumption," in *Proc. IEEE PES Conf. on Innovative Smart Grid Technologies (ISGT)*, Feb. 2014, pp. 1–5.

[104] A. Kavousian, R. Rajagopal, and M. Fischer, "Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior," *Energy*, vol. 55, no. 15, pp. 184–194, Jun. 2013.

[105] Z. Yu, B. C. Fung, F. Haghighat, H. Yoshino, and E. Morofsky, "A systematic procedure to study the influence of occupant behavior on building energy consumption," *Energy*, vol. 43, no. 6, pp. 1409–1417, Jun. 2011.

[106] L. Wu, G. Kaiser, D. Solomon, R. Winter, A. Boulanger, and R. Anderson, "Improving efficiency and reliability of building systems using machine learning and automated online evaluation," in *Proc. IEEE Long Island Systems, Applications and Technology Conf. (LISAT)*, May 2012, pp. 1–6.

[107] S. Li, K. Deng, and M. Zhou, "Social incentive policies to engage commercial building occupants in demand response," in *Proc. IEEE Intl. Conf. on Automation Science and Engineering (CASE)*, Aug. 2014, pp. 407–412.

[108] B. J. Johnson, M. R. Starke, O. A. Abdelaziz, R. K. Jackson, and L. M. Tolbert, "A MATLAB based occupant driven dynamic model for predicting residential

power demand," in *Proc. IEEE PES Trans. Dist. Conf. and Exp.*, Apr. 2014, pp. 1–5.

[109] K. Anderson and S. Lee, "Modeling occupant energy use interventions in evolving social networks," in *Proc. Winter Simulation Conf. (WSC)*, Dec. 2013, pp. 3051–3058.

[110] Y. Yang, Q. S. Jia, and X. Guan, "Improving the prediction accuracy of building energy consumption using location of occupant: A case study," in *Proc. IEEE Intel. Conf. on Industrial Technology (ICIT)*, Mar. 2016, pp. 1550–1555.

[111] Y. G. Yohanis, J. D. Mondol, A. Wright, and B. Norton, "Real-life energy use in the UK: How occupancy and dwelling characteristics affect domestic electricity use," *Energy and Buildings*, vol. 40, no. 6, pp. 1053–1059, 2008.

[112] A. C. Menezes, A. Cripps, D. Bouchlaghem, and R. Buswell, "Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap," *Applied Energy*, vol. 97, pp. 355–364, Sept. 2012.

[113] J. Fiksel, "Sustainabilty and resilience: toward a systems approach," *IEEE Engineering Management Review*, vol. 35, no. 3, pp. 1–5, Aug. 2007.

[114] P. S. N. Rao and R. Deekshit, "Energy loss estimation in distribution feeders," *IEEE Trans. Power Del.*, vol. 21, no. 3, pp. 1092–1100, Jul. 2006.

[115] R. Subbiah, K. Lum, A. Marathe, and M. Marathe, "A high resolution energy demand model for commercial buildings," in *Proc. Intl. ETG-Congress Security in Critical Infrastructures Today*, Nov. 2013, pp. 1–6.

[116] G. Peschiera, J. E. Taylor, and J. A. Siegel, "Response-relapse patterns of building occupant electricity consumption following exposure to personal, contextualized and occupant peer network utilization data," *Energy and Buildings*, vol. 42, no. 8, pp. 1329–1336, Aug. 2010.

[117] Y. Tang, S. Zhao, C.-W. Ten, and K. Zhang, "Enhancement of distribution load modeling using statistical hybrid regression," in *Proc. IEEE PES Conf. on Innovative Smart Grid Technologies (ISGT)*, Apr. 2017, pp. 1–5.

[118] X. Zhou and H. S. Mahmassani, "A structural state space model for real-time traffic origindestination demand estimation and prediction in a day-to-day learning framework," *Transportation Research Part B: Methodological*, vol. 41, no. 8, pp. 823–840, Oct. 2007.

[119] Z. Ghadialy. (2007, Mar.) Assisted Global Positioning System. [Online]. Available: http://www.3g4g.co.uk/Faq/agps.html.

[120] I. Eusgeld, C. Nan, and S. Dietz, ""System-of-systems" approach for interdependent critical infrastructures," *Reliabilty Eng. and Syst. Safety*, vol. 96, no. 6, pp. 679–686, Jun. 2011.

[121] (2012, Dec.) GridLAB-D. U.S. Department of Energy (DOE) at Pacific

Northwest National Laboratory (PNNL). [Online]. Available: http://www.gridlabd.org.

[122] (2012, Sept.) Occupantload. U.S. Department of Energy (DOE) at Pacific Northwest National Laboratory (PNNL). [Online]. Available: http://gridlab-d.shoutwiki.com/wiki/Occupantload.

[123] R. Jennrich, *An Introduction to Computational Statistics*, 1st ed. Prentice-Hall, Inc., 1995.

[124] O. Feldman, "The GEH measure and quality of the highway assignment models," in *Proc. European Transport Conf.*, Oct. 2012, pp. 1–18.

[125] (2012, Oct.) Beginner's Guide to GridLAB-D. U.S. Department of Energy (DOE) at Pacific Northwest National Laboratory (PNNL). [Online]. Available: http://gridlab-d.shoutwiki.com/wiki/Beginner%27s_Guide_to_GridLAB-D.

[126] (2013, Nov.) IEEE 13-node with houses. GitHub. [Online]. Available: https://github.com/GridOPTICS/FNCS-gridlab-d/blob/master/models/IEEE_13_Node_With_Houses.glm.

[127] (2017, Jul.) ArcGIS. Esri. [Online]. Available: https://www.arcgis.com/features/index.html.

[128] (2017, Apr.) Michigan Tech Smart Grid Data Download. Yachen Tang. [Online]. Available: https://tangdean1214.wixsite.com/occupancyconsumption.

[129] C. Cody, V. Ford, and A. Siraj, "Decision tree learning for fraud detection in consumer energy consumption," in *Proc. IEEE 14th Intl. Conf. on Machine Learning and Applications*, Dec. 2015, pp. 1175–1179.

[130] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity Theft Detection in AMI Using Customers' Consumption Patterns," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan. 2016.

[131] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *IEEE Security Privacy*, vol. 7, no. 3, pp. 75–77, Jun. 2009.

[132] C. Fu, C. Wang, C. Miller, and Y. Wang, "Characterization and distribution model for fuel costs in electricity market," in *Energy and Sustainability Conf. (IESC)*, Oct. 2017, pp. 1–6.

[133] R. Czechowski and A. M. Kosek, "The most frequent energy theft techniques and hazards in present power energy consumption," in *Joint Workshop on CPSR-SG*, Apr. 2016, pp. 1–7.

[134] S. Sahoo, D. Nikovski, T. Muso, and K. Tsuru, "Electricity theft detection using smart meter data," in *IEEE ISGT*, Feb. 2015, pp. 1–5.

[135] Y. Zhou, X. Chen, A. Y. Zomaya, L. Wang, and S. Hu, "A dynamic programming algorithm for leveraging probabilistic detection of energy theft in smart home," *IEEE Trans. Emerging Topics in Computing*, vol. 3, no. 4, pp. 502–513, Dec. 2015.

[136] S. A. Salinas and P. Li, "Privacy-preserving energy theft detection in microgrids: A state estimation approach," *IEEE Trans. Power Syst.*, vol. 31, no. 2, pp. 883–894, Mar. 2016.

[137] B. Coma-Puig, J. Carmona, R. Gavald, S. Alcoverro, and V. Martin, "Fraud detection in energy consumption: A supervised approach," in *IEEE Intl. Conf. on DSAA*, Oct. 2016, pp. 120–129.

[138] R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, and X. S. Shen, "Energy-theft detection issues for advanced metering infrastructure in smart grid," *Tsinghua Sci. and Tech.*, vol. 19, no. 2, pp. 105–120, Apr. 2014.

[139] J. E. Cabral, J. O. P. Pinto, and A. M. A. C. Pinto, "Fraud detection system for high and low voltage electricity consumers based on data mining," in *IEEE Power Energy Society General Meeting*, Jul. 2009, pp. 1–5.

[140] J. E. Cabral and E. M. Gontijo, "Fraud detection in electrical energy consumers using rough sets," in *IEEE Intl. Conf. on Syst., Man and Cybernetics*, Oct. 2004, pp. 3625–3629.

[141] A. Augugliaro, L. Dusonchet, M. G. Ippolito, and E. R. Sanseverino, "Minimum losses reconfiguration of mv distribution networks through local control of tie-switches," *IEEE Trans. Power Del.*, vol. 18, no. 3, pp. 762–771, Jul. 2003.

[142] D. Jiang and R. Baldick, "Optimal electric distribution system switch reconfiguration and capacitor control," *IEEE Trans. Power Syst.*, vol. 11, no. 2, pp. 890–897, May 1996.

[143] A. Abiri-Jahromi, M. Fotuhi-Firuzabad, M. Parvania, and M. Mosleh, "Optimized sectionalizing switch placement strategy in distribution systems," *IEEE Trans. Power Del.*, vol. 27, no. 1, pp. 362–369, Jul. 2012.

[144] E. D. Tuglie, M. L. Scala, G. Patrono, P. Pugliese, and F. Torelli, "An optimal strategy for switching devices allocation in radial distribution network," in *IEEE Africon. 7th Africon Conf. in Africa*, Sept. 2004, pp. 683–689.

[145] J. Li, X. Y. Ma, C. C. Liu, and K. P. Schneider, "Distribution system restoration with microgrids using spanning tree search," *IEEE Trans. Power Syst.*, vol. 29, no. 6, pp. 3021–3029, Nov. 2014.

[146] L. S. de Assis, J. F. V. Gonzlez, F. L. Usberti, C. Lyra, C. Cavellucci, and F. J. V. Zuben, "Switch allocation problems in power distribution systems," *IEEE Trans. Power Syst.*, vol. 30, no. 1, pp. 246–253, Jan. 2015.

[147] A. Mendes, N. Boland, P. Guiney, and C. Riveros, "Switch and tap-changer reconfiguration of distribution networks using evolutionary algorithms," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 85–92, Feb. 2013.

[148] A. Heidari, V. G. Agelidis, M. Kia, J. Pou, J. Aghaei, M. Shafie-Khah, and J. P. S. Catalo, "Reliability optimization of automated distribution networks

with probability customer interruption cost model in the presence of dg units," *IEEE Trans. Smart Grid*, vol. 8, no. 1, pp. 305–315, Jan. 2017.

[149] Y. Tang, C. Ten, and L. E. Brown, "Switching reconfiguration of fraud detection within an electrical distribution network," in *Resilience Week (RWS)*, Sept 2017, pp. 206–212.

[150] Z. W. X. C. C. Yuan, G. Liu and M. S. Illindala, "Economic power capacity design of distributed energy resources for reliable community microgrids," *Energy Procedia*, vol. 142, p. 25612567, Dec. 2017.

[151] S. N. Duanka Janei, Ante Milievi and N. Trinajsti. (2009, Sept.) Graph theoretical matrices in chemistry. [Online]. Available: http://www.sicmm.org/~FAMNIT-knjiga/wwwANG/index.htm.

[152] M. Technology. (2017, May.) Pairing function. [Online]. Available: http://mathworld.wolfram.com/PairingFunction.html.

[153] Ondrej. (2011, Jul.) Graph adjacency matrix to incidence matrix. [Online]. Available: https://www.mathworks.com/matlabcentral/fileexchange/24661-graph-adjacency-matrix-to-incidence-matrix?focused=5191716&tab=function.

[154] (2011, Jul.) Convert adjacency matrix to an incidence matrix. [Online]. Available: http://strategic.mit.edu/docs/matlab_networks/adj2inc.m.

[155] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE Trans. Power Del.*, vol. 25, no. 2, pp. 1162–1171, Apr. 2010.

[156] M. Davis. (2009) SmartGrid Device Security: Adventures in a new medium. Black Hat. [Online]. Available: http://www.blackhat.com/presentations/bh-usa-09/MDAVIS/BHUSA09-Davis-AMI-SLIDES.pdf.

[157] D. M. Nicol, "Hacking the lights out: The computer virus threat to the electrical grid," *Scientific American*, vol. 305, no. 1, pp. 70–75, Jul. 2011.

[158] J. Pollet. (2010) Electricity for Free? The Dirty Underbelly of SCADA and Smart Meters. Red Tiger Security. [Online]. Available: https://media.blackhat.com/bh-us-10/whitepapers/Pollet_Cummins/BlackHat-USA-2010-Pollet-Cummings-RTS-Electricity-for-Free-wp.pdf.

[159] K. Cho, M. Jo, T. Kwon, H. Chen, and D. Lee, "Classification and experimental analysis for clone detection approaches in wireless sensor networks," *IEEE Systems Journal*, vol. 7, no. 1, pp. 26–35, Mar. 2013.

[160] M. Demirbas and Y. Song, "An RSSI-based scheme for sybil attack detection in wireless sensor networks," in *Proc. WoWMoM 2006*, Jun. 2006, pp. 566–570.

[161] (2012) smarter protection for the smart grid. McAfee. [Online]. Available: https://www.ccn-cert.cni.es/publico/InfraestructurasCriticaspublico/rp-smarter-protection-smart-grid.pdf.

[162] V. Ford, A. Siraj, and W. Eberle, "Smart grid energy fraud detection using artificial neural networks," in *Proc. IEEE 8th Intl. Symposium on Service Oriented System Engineering*, Dec. 2014, pp. 1–6.

[163] E. Choo, Y. Park, and H. Siyamwala, "Identifying malicious metering data in advanced metering infrastructure," in *Proc. IEEE Symposium on Computational Intelligence Applications in Smart Grid (CIASG)*, Apr. 2014, pp. 490–495.

[164] Y. Park, D. M. Nicol, H. Zhu, and C. W. Lee, "Prevention of malware propagation in AMI," in *Proc. IEEE Intl. Conf. on Smart Grid Communications (SmartGridComm)*, Oct. 2013, pp. 474–479.

[165] H. Sedjelmaci and S. M. Senouci, "Smart grid security: A new approach to detect intruders in a smart grid neighborhood area network," in *Proc. Intl. Conf. on Wireless Networks and Mobile Communications (WINCOM)*, Oct. 2016, pp. 6–11.

[166] N. Beigi-Mohammadi, J. Mii, H. Khazaei, and V. B. Mii, "An intrusion detection system for smart grid neighborhood area network," in *Proc. IEEE Intl. Conf. on Communications (ICC)*, Jun. 2014, pp. 4125–4130.

[167] X. Xia, W. Liang, Y. Xiao, and M. Zheng, "BCGI: A fast approach to detect malicious meters in neighborhood area smart grid," in *Proc. IEEE Intl. Conf. on Communications (ICC)*, Jun. 2015, pp. 7228–7233.

[168] Z. Xiao, Y. Xiao, and D. H.-C. Du, "Exploring malicious meter inspection in neighborhood area smart grids," *IEEE Trans. Smart Grid*, vol. 4, no. 1, pp. 214–226, 2013.

[169] X. Xia, W. Liang, Y. Xiao, M. Zheng, and Z. Xiao, "A difference-comparison-based approach for malicious meter inspection in neighborhood area smart grids," in *Proc. IEEE Intl. Conf. on Communications (ICC)*, Jun. 2015, pp. 802–807.

[170] C. Yuan and M. Illindala, "Economic sizing of distributed energy resources for reliable community microgrids," in *IEEE Power and Energy Society General Meeting*, Jul. 2017, pp. 1–5.

[171] C. Yuan, M. Illindala, M. Haj-ahmed, and A. Khalsa, "Distributed energy resource planning for microgrids in the united states," in *IEEE Industry Applications Society Annual Meeting*, Oct. 2015, pp. 18–22.

[172] C. Yuan, M. Illindala, and A. Khalsa, "Co-optimization scheme for energy resource planning in community microgrids," *IEEE Trans. Sustainable Energy*, vol. 8, no. 4, pp. 1351–1360, Oct. 2017.

[173] Z.Yang and C.-W. Ten, "Assessment of hypothesized substation cyberattack using linearized power flow approach," in *ISGT*, Apr. 2017, pp. 1–5.

[174] A. H. Nizar, Z. Y. Dong, and Y. Wang, "Power utility nontechnical loss analysis

with extreme learning machine method," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 946–955, Aug. 2008.

[175] Y. Guo, C. W. Ten, and P. Jirutitijaroen, "Online data validation for distribution operations against cybertampering," *IEEE Trans. Power Syst.*, vol. 29, no. 2, pp. 550–560, Mar. 2014.

[176] S. Tonyali. (2018, Jul.) Security and privacy concerns in smart metering: The cyber-physical aspect. IEEE Smart Grid. [Online]. Available: https://smartgrid.ieee.org/newsletters/july-2018/security-and-privacy-concerns-in-smart-metering-the-cyber-physical-aspect.

[177] S.-C. Yip, W.-N. Tan, C.-K. Tan, M.-T. Gan, and K. Wong, "An anomaly detection framework for identifying energy theft and defective meters in smart grids," *Intl. Journal of Electrical Power and Energy Systems*, vol. 101, pp. 189–203, Jul. 2018.

[178] Y. Guo, C. W. Ten, S. Hu, and W. W. Weaver, "Preventive maintenance for advanced metering infrastructure against malware propagation," *IEEE Trans. Smart Grid*, vol. 7, no. 3, pp. 1314–1328, May 2016.

[179] C.-W. Ten, K. Yamashita, Z. Yang, A. V. Vasilakos, and A. Ginter, "Impact assessment of hypothesized cyberattacks on interconnected bulk power systems," *IEEE Trans. Smart Grid*, vol. 9, no. 3, pp. 4405–4425, Sept. 2018.

[180] M. Gander, C. Sauerwein, and R. Breu, "Assessing real-time malware threats,"

in *IEEE Intl. Conf. on Software Quality, Reliability and Security - Companion*, Aug. 2015, pp. 6–13.

[181] T. Barabosch and E. Gerhards-Padilla, "Host-based code injection attacks: A popular technique used by malware," in *9th Intl. Conf. on Malicious and Unwanted Software: The Americas (MALWARE)*, Oct. 2014, pp. 8–17.

[182] Z.Yang and C.-W. Ten, "Cyber-induced risk modeling for microprocessor-based relays in substations," in *ISGT Asia*, May 2018, pp. 856–861.

[183] S. Paul and Z. Ni, "Vulnerability analysis for simultaneous attack in smart grid security," in *IEEE Power Energy Society Innovative Smart Grid Technologies Conf. (ISGT)*, Apr. 2017, pp. 1–5.

[184] (2016) Analysis of the Cyber Attack on the Ukrainian Power Grid. E-ISAC. [Online]. Available: https://www.nerc.com/pa/CI/ESISAC/Documents/E-ISAC_SANS_Ukraine_DUC_18Mar2016.pdf.

[185] V. Dehalwar, A. Kalam, M. L. Kolhe, and A. Zayegh, "Review of detection, assessment and mitigation of security risk in smart grid," in *2nd Intl. Conf. on Power and Renewable Energy (ICPRE)*, Sept. 2017, pp. 1077–1081.

[186] F. Grbert, A. Sadeghi, and M. Winandy, "Software distribution as a malware infection vector," in *Intl. Conf. for Internet Technology and Secured Transactions, (ICITST)*, Nov. 2009, pp. 1–6.

[187] J. Zhang, S. Saha, G. Gu, S. Lee, and M. Mellia, "Systematic mining of associated server herds for malware campaign discovery," in *IEEE 35th Intl. Conf. on Distributed Computing Systems*, Jun. 2015, pp. 630–641.

[188] E. Filiol and S. Josse, "New trends in security evaluation of bayesian network-based malware detection models," in *45th Hawaii Intl. Conf. on System Sciences*, Jan. 2012, pp. 5574–5583.

[189] K. C. Budka, J. G. Deshpande, T. L. Doumi, M. Madden, and T. Mew, "Communication network architecture and design principles for smart grids," *Bell Labs Technical Journal*, vol. 15, no. 2, pp. 205–227, Sept. 2010.

[190] Z.Yang, C.-W. Ten, and A. Ginter, "Extended enumeration of hypothesized substations outages incorporating overload implication," *IEEE Trans. Smart Grid*, vol. 9, no. 6, pp. 6929–6938, Nov. 2018.